Model-Based Gait and Action Recognition Using Kinect

by

Faisal Ahmed

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN COMPUTER SCIENCE

CALGARY, ALBERTA
July, 2016

© Faisal Ahmed 2016
Abstract

Being the very first in the category of low-cost consumer-level depth sensors, the recent release of Microsoft Kinect has opened the door to a new generation of computer vision and biometric security applications. This thesis focuses on designing new methodologies for Kinect-based gait and action recognition systems that utilize the Kinect 3D virtual skeleton to construct effective and robust motion representations. The proposed gait recognition method focuses on designing a feature descriptor that can capture person-specific distinct motion patterns, caused by the influence of human physiology and behavioral traits. On the other hand, the proposed action recognition method involves constructing a person-independent feature descriptor that can suppress person-specific motion traits while highlighting a more generic and high level description of action-specific skeletal joint movements. Extensive experiments with three recently released public benchmark databases demonstrate the effectiveness of the proposed methodologies, compared against state-of-the-art gait and action recognition methods.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>iii</td>
</tr>
<tr>
<td>List of Figures and Illustrations</td>
<td>v</td>
</tr>
<tr>
<td>List of Symbols, Abbreviations and Nomenclature</td>
<td>vii</td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Problem Statement</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Objectives</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Contributions</td>
<td>6</td>
</tr>
<tr>
<td>1.5 Thesis Outline</td>
<td>8</td>
</tr>
<tr>
<td>2 LITERATURE REVIEW</td>
<td>10</td>
</tr>
<tr>
<td>2.1 Background</td>
<td>10</td>
</tr>
<tr>
<td>2.2 Overview of Gait and Action Recognition Methods</td>
<td>12</td>
</tr>
<tr>
<td>2.3 Vision-Based Gait Recognition Systems</td>
<td>15</td>
</tr>
<tr>
<td>2.3.1 Model-Based Gait Recognition</td>
<td>16</td>
</tr>
<tr>
<td>2.3.2 Appearance-Based Gait Recognition</td>
<td>17</td>
</tr>
<tr>
<td>2.3.3 Gait Recognition Using Kinect</td>
<td>19</td>
</tr>
<tr>
<td>2.4 Vision-Based Human Action Recognition Systems</td>
<td>20</td>
</tr>
<tr>
<td>2.4.1 Single and Multi-Camera Human Action Recognition</td>
<td>21</td>
</tr>
<tr>
<td>2.4.2 Human Action Recognition Using Kinect</td>
<td>22</td>
</tr>
<tr>
<td>2.5 Concluding Remarks</td>
<td>23</td>
</tr>
<tr>
<td>3 PROPOSED GAIT RECOGNITION METHOD</td>
<td>25</td>
</tr>
<tr>
<td>3.1 Overview</td>
<td>25</td>
</tr>
<tr>
<td>3.2 Gait Cycle Detection</td>
<td>26</td>
</tr>
<tr>
<td>3.3 Gait Feature Representation</td>
<td>28</td>
</tr>
<tr>
<td>3.4 Relevant Joint-Pair Selection</td>
<td>30</td>
</tr>
<tr>
<td>3.5 DTW-Based Classifier</td>
<td>33</td>
</tr>
<tr>
<td>3.6 Score Level Fusion of JRD and JRA</td>
<td>36</td>
</tr>
<tr>
<td>3.7 Summary</td>
<td>36</td>
</tr>
<tr>
<td>4 PROPOSED ACTION RECOGNITION METHOD</td>
<td>38</td>
</tr>
<tr>
<td>4.1 Overview</td>
<td>38</td>
</tr>
<tr>
<td>4.2 Joint-Triplet Motion Image (JTMI)</td>
<td>40</td>
</tr>
<tr>
<td>4.3 Local Binary Pattern (LBP) Texture Descriptor</td>
<td>42</td>
</tr>
<tr>
<td>4.4 Fisher Linear Discriminant Analysis</td>
<td>45</td>
</tr>
<tr>
<td>4.5 Summary</td>
<td>47</td>
</tr>
<tr>
<td>5 EXPERIMENTAL RESULTS</td>
<td>49</td>
</tr>
<tr>
<td>5.1 Experimental Results for Gait Recognition</td>
<td>49</td>
</tr>
<tr>
<td>5.1.1 Experimental Setup and Dataset Description</td>
<td>50</td>
</tr>
<tr>
<td>5.1.2 Experimental Results for JRD-Based Gait Recognition</td>
<td>52</td>
</tr>
<tr>
<td>5.1.3 Experimental Results for JRA-Based Gait Recognition</td>
<td>56</td>
</tr>
<tr>
<td>5.1.4 Fusion of JRD and JRA-Based Gait Recognition</td>
<td>61</td>
</tr>
<tr>
<td>5.2 Experimental Results for Action Recognition</td>
<td>65</td>
</tr>
</tbody>
</table>
## List of Figures and Illustrations

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Different data streams obtained from the Kinect sensor.</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>3D skeletal joints tracked by the Kinect sensor. A hierarchy of the bones</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>constructed based on the tracked skeletal joints is also shown.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Each bone is specified as a parent and a child node in this hierarchy.</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>Categories of gait and action recognition methods.</td>
<td>14</td>
</tr>
<tr>
<td>2.3</td>
<td>Overview of a generic vision-based gait recognition system.</td>
<td>16</td>
</tr>
<tr>
<td>2.4</td>
<td>Overview of a generic vision-based action recognition system.</td>
<td>21</td>
</tr>
<tr>
<td>3.1</td>
<td>Overview of the proposed gait recognition method.</td>
<td>27</td>
</tr>
<tr>
<td>3.2</td>
<td>Detection of a complete gait cycle based on the Ankle Left and Ankle Right</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>joint distance.</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>JRD sequences for different joint-pairs and persons computed over a complete</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>gait cycle. Here, the x-axis represents frames and y-axis represents</td>
<td></td>
</tr>
<tr>
<td></td>
<td>normalized JRD values. It can be observed that (a) some JRD sequences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>have high variations, while (b) the others have a flat pattern.</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Proposed score-level fusion of JRD and JRA features based on DTW classifier</td>
<td>37</td>
</tr>
<tr>
<td>4.1</td>
<td>Overview of the proposed action recognition method.</td>
<td>39</td>
</tr>
<tr>
<td>4.2</td>
<td>Construction of the 2D joint-triplet motion image.</td>
<td>43</td>
</tr>
<tr>
<td>4.3</td>
<td>(a) The basic LBP encoding scheme, (b) Overview of the LBP feature descriptor</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>construction process.</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Sample skeletal gait sequences of a participant from the UPCV gait dataset</td>
<td>51</td>
</tr>
<tr>
<td>5.2</td>
<td>Heat map of the generalized joint-pair relevance matrix computed from the</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>whole UPCV gait database. Here, each cell (i,j) represents the normalized</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean flatness of the JRD sequences of joint-pair {i, j}.</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Number of relevant joint-pairs for different threshold (t) values. For a</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>given t, only the JRD sequences that has the normalized mean flatness equal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>or less than t is selected for the final gait feature representation.</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>Selected joint-pairs based on the normalized mean flatness measure</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>(obtained from the generalized JPR matrix) for different threshold values.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Here, the cells with value 0 (dark blue) corresponds to the joint-pairs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>which are not used in the final feature representation.</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>Rank 1 recognition performance of the proposed JRD-based gait recognition</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>for different threshold values.</td>
<td></td>
</tr>
<tr>
<td>5.6</td>
<td>CMC curve for JRD-based gait recognition using different threshold (t) values.</td>
<td>57</td>
</tr>
</tbody>
</table>
5.7 Heat map of the generalized joint-pair relevance matrix computed from the whole UPCV gait database. Here, each cell \((i,j)\) represents the normalized mean flatness of the JRA sequences of joint-pair \(\{i,j\}\).

5.8 Number of relevant joint-pairs for different threshold \((t)\) values. For a given \(t\), only the JRA sequences that has the normalized mean flatness equal or less than \(t\) is selected for the final gait feature representation.

5.9 Selected joint-pairs based on the normalized mean flatness measure of JRA sequences (obtained from the generalized JPR matrix) for different threshold values. Here, the cells with value 0 (dark blue) corresponds to the joint-pairs which are not used in the final feature representation.

5.10 Rank 1 recognition performance of the proposed JRA-based gait recognition for different threshold \((t)\) values.

5.11 CMC curve for JRA-based gait recognition using different threshold \((t)\) values.

5.12 CMC curve for different gait recognition methods.

5.13 Sample RGB and depth frames from the UTKinect Action Dataset for different actions [2]. It can be observed that the actions have been recorded from different views.

5.14 Sample RGB and depth frames from the KARD dataset for different activities [3].

5.15 Recognition performance for the UTKinect dataset for different number of sub-regions into which the LBP encoded JTMI image is partitioned. Only the LBP histogram features are used for the classification task.

5.16 Recognition performance for the UTKinect dataset for different normalization of JRA sequences. Only the LBP histogram features are used for the classification task.

5.17 Recognition performance (represented as CMS score (%)) of different methods for the UTKinect dataset.

5.18 Confusion matrix of the proposed method for the UTKinect Dataset. Rows represent true class and columns represent recognition rate (%).

5.19 Recognition performance (represented as CMS score (%)) of different methods for the KARD dataset.
## List of Symbols, Abbreviations and Nomenclature

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AME</td>
<td>Accumulated Motion Energy</td>
</tr>
<tr>
<td>ATM</td>
<td>Automated Teller Machine</td>
</tr>
<tr>
<td>CMC</td>
<td>Cumulative Match Characteristics</td>
</tr>
<tr>
<td>CMS</td>
<td>Cumulative Match Score</td>
</tr>
<tr>
<td>CMOS</td>
<td>Complementary Metal-Oxide Semiconductor</td>
</tr>
<tr>
<td>CRF</td>
<td>Conditional Random Field</td>
</tr>
<tr>
<td>DGHEI</td>
<td>Depth Gradient Histogram Energy Image</td>
</tr>
<tr>
<td>DTW</td>
<td>Dynamic Time Warping</td>
</tr>
<tr>
<td>DWT</td>
<td>Discrete Wavelet Transformation</td>
</tr>
<tr>
<td>EMG</td>
<td>Electromyography</td>
</tr>
<tr>
<td>ETS</td>
<td>Electromagnetic Tracking System</td>
</tr>
<tr>
<td>FDEI</td>
<td>Frame Difference Energy Image</td>
</tr>
<tr>
<td>FEI</td>
<td>Foot Energy Image</td>
</tr>
<tr>
<td>FLD</td>
<td>Fisher Linear Discriminant</td>
</tr>
<tr>
<td>GEI</td>
<td>Gait Energy Image</td>
</tr>
<tr>
<td>GEV</td>
<td>Gait Energy Volume</td>
</tr>
<tr>
<td>HEI</td>
<td>Head Energy Image</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>HOJ3D</td>
<td>Histogram of 3D Joints</td>
</tr>
<tr>
<td>IR</td>
<td>Infrared</td>
</tr>
<tr>
<td>JRA</td>
<td>Joint Relative Angle</td>
</tr>
<tr>
<td>JRD</td>
<td>Joint Relative Distance</td>
</tr>
<tr>
<td>JTMI</td>
<td>Joint-Triplet Motion Image</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>KARD</td>
<td>Kinect Action Recognition Dataset</td>
</tr>
<tr>
<td>k-NN</td>
<td>k-Nearest Neighbor</td>
</tr>
<tr>
<td>LBP</td>
<td>Local Binary Pattern</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>LDM</td>
<td>Layered Deformable Model</td>
</tr>
<tr>
<td>MDI</td>
<td>Multiple Discriminant Analysis</td>
</tr>
<tr>
<td>MEI</td>
<td>Motion Energy Image</td>
</tr>
<tr>
<td>MHI</td>
<td>Motion History Image</td>
</tr>
<tr>
<td>MOCAP</td>
<td>Motion Capture</td>
</tr>
<tr>
<td>NBNN</td>
<td>Naive-Bayes-Nearest-Neighbor</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PDV</td>
<td>Pose Depth Volume</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
</tr>
<tr>
<td>SMIJ</td>
<td>Sequence of the Most Informative Joints</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>UPCV</td>
<td>University of Patras, Computer Vision Group</td>
</tr>
</tbody>
</table>
Chapter 1

INTRODUCTION

1.1 Motivation

The primary objective of my thesis is to develop effective gait and action recognition systems based on the recently released Microsoft Kinect sensor. Vision-based human gait and action recognition are two of the most fundamental research problems in the areas of biometric security and computer vision, with potential applications in security and surveillance, assisted living and healthcare, motion and video retrieval, intelligent systems, etc. Biometric gait recognition focuses on identifying a person based on the distinctiveness in the movement patterns of different body limbs during walking. On the other hand, the objective in vision-based action recognition is to understand actions and activities performed by an individual or a group. Traditional approaches to human gait and action recognition involve video-based analysis of motion patterns of different body limbs. With the advancements in sensor technology, modern day consumer-level video cameras have become inexpensive, which has facilitated extensive use of these cameras in gait and action recognition [4, 5]. However, extracting motion data from 2D intensity image sequences requires extensive pre-processing, which involves background modeling and subtraction, and extraction of human shape or silhouette motion. Hence, the involved computational cost makes video camera-based gait and action recognition difficult to accommodate in many real-world scenarios. On the other hand, although sophisticated multi-camera motion capture systems can extract reliable motion data in real-time, the cost involved makes widespread deployment of such systems not feasible [6].

The recent release of the Microsoft Kinect has changed the research landscape,
providing scientists with an inexpensive consumer-level depth sensor that can capture human motion data in real-time [7]. Here, one of the main differences from traditional video-based approaches is that the depth and skeleton data are directly available from the sensor, which renders the time-consuming video pre-processing tasks unnecessary [7]. The complementary nature of the multimodal data streams (RGB, depth, and skeleton) provided by Kinect has the potential to facilitate computers to attain a realistic perception of users, objects, and actions in a 3D environment. As a result, the Kinect sensor has attracted a significant attention from the computer vision and biometrics research community to develop new generations of real-time computer vision and security-related applications. In this context, biometric gait and action recognition using Kinect has great potential to replace modern day video-based security and surveillance systems as well as home monitoring systems for assisted living.

In my thesis, I consider using the Kinect skeletal joint coordinate data for effective 3D gait and action recognition. The primary research questions addressed in this thesis are the followings:

1. Can the Kinect skeletal motion data be utilized to construct an effective biometric gait recognition system?
2. Can the Kinect skeletal motion data be utilized to construct an effective human action recognition system?
3. Can we construct view and scale invariant gait and action feature representations based on the Kinect skeletal motion data that can effectively capture the underlying motion patterns of different body limbs?

1.2 Problem Statement

Secure real-time user authentication and access control management are crucial for a wide variety of systems and services, such as secure access to facilities, cell phones,
automated teller machines (ATM), computers, etc. Traditional approaches to the problem involve use of certain tokens, such as identification (ID) card or password verification. However, these solutions have some drawbacks. For example, ID cards can be lost, stolen, or forged, while passwords can be compromised or forgotten [8]. As a result, these approaches are vulnerable to forgery and unable to provide sufficient security [9]. In recent years, rapid development of biometric technologies has opened the door to a new class of fast and reliable identity management solutions [10], which are being actively researched and deployed in both corporate and academic settings [11, 12, 13]. A biometric system can be defined as a pattern-recognition system that can recognize individuals based on the characteristics of their physiology or behavior [8]. Human biometric traits can roughly be divided into two categories: physiological and behavioral. Physiological biometric systems utilize certain physical characteristics, such as face, iris, ear, fingerprints, palmprints, etc. for individual recognition. On the other hand, behavioral biometric systems rely on human behavior-mediated activities, such as gait, voice, handwriting, signature, etc. In most cases, a biometric system requires direct participation or cooperation of the person (in the form of physical contact or pose) being recognized [7, 14]. However, there are only a few biometrics that can be recognized unobtrusively without any cooperation of the user. Gait is one such behavioral biometric which can be defined as the movement patterns of certain
body limbs and their interactions with the surrounding environment during walking \[15\] \[16\]. It is a complex and dynamic behavioral trait, which makes it difficult to disguise someone’s own gait or imitate some other person’s gait \[17\]. This characteristic makes gait recognition particularly useful in scenarios where other biometric traits are obscured (often intentionally, such as a crime scene) or user cooperation is not intended (such as surveillance in public places like airports, bus stations, etc.) \[17\]. In addition, gait analysis can potentially be utilized in virtual and augmented reality, motion and video retrieval \[18\], 3D human body modeling and animation \[19\], \[20\], healthcare \[21\], etc.

While gait recognition focuses on a particular human action (walking) to extract distinctive biometric signature, the objective of human action recognition is to understand an action instance performed by an individual or a group, often under varying conditions \[22\]. Human action recognition is a widely-studied area in computer vision due to its potential applicability in surveillance and security \[23\], assisted living and healthcare \[24\], motion retrieval \[25\], and human-computer interfaces \[26\], \[27\]. Human actions typically involve movements of body limbs that can be defined at various levels of abstraction \[23\]. In this thesis, I adopt the hierarchy proposed by Moeslund et al. \[6\], which comprises action primitive, action, and activity. An action primitive can be defined as an atomic movement of a limb (e.g. right leg forward). On the other hand, an action is composed of one or multiple atomic limb movements (action primitives), often in a cyclic pattern (e.g. running, walking, jumping, etc.). Lastly, an activity comprises a number of actions, typically performed in a specific sequence (e.g. cooking, playing football, etc.).

Generic gait and action recognition systems typically comprise three basic components: i) a sensor or camera to collect the data, ii) a feature representation, and iii) a classifier. The feature representation describes the characteristics of the captured
gait or motion data (obtained from the sensor), which is then used by the classifier to recognize the person or action. However, even using the best classifier will result in poor recognition performance, if provided with features having low discriminating ability or inadequate information. Hence, designing an effective and discriminative feature representation is one of the main challenges in gait and action recognition. An effective and discriminative feature descriptor can be characterized as i) having high inter-class variations and low intra-class variations, ii) providing robustness in uncontrolled environment under the presence of view and scale changes, and iii) having a low-dimensional feature space to facilitate low computational cost \[28, 29\]. Since gait and action recognition using Kinect is an emerging area of research, most of the existing studies have been conducted under simplifying assumptions, such as constant walking or movement speed, fixed distance and viewpoint, relatively smaller datasets, etc. However, in order to deploy gait and action recognition systems in real-world settings, constructing effective feature representations that can work in unconstrained environment is an important task. In summary, the challenges I aim to tackle in this thesis are: i) finding a set of discriminative features that are robust against view and scale changes as well as the speed of walking or performing an action, and ii) utilizing these features in designing effective gait and action recognition systems that can attain high recognition performance in uncontrolled environment.

1.3 Objectives

The primary goal of my thesis is to develop effective methodologies for biometric gait and action recognition using the Kinect sensor. In particular, I aim to investigate how the 3D skeletal motion data obtained from the Kinect can effectively be utilized to design real-world gait and action recognition systems. The developed methodologies should satisfy the following criteria:
1. The constructed feature descriptors should be view and scale-invariant and should effectively represent the human gait and action-specific motion patterns in a robust manner.

2. To ensure low computational cost, the developed methodologies should incorporate effective feature selection techniques to obtain highly discriminating feature representations in a low-dimensional space.

3. The developed methodologies should not require individuals to walk only to a specific direction (fronto-normal or fronto-parallel) or always maintain a fixed distance from the camera while performing an action.

4. The developed methodologies should be robust against variations in the speed of walking or performing particular actions resulting in variable length video sequences.

5. Lastly, I investigate whether fusing disparate feature representations can potentially improve the gait recognition performance.

1.4 Contributions

In this thesis, I present new methodology for gait and action recognition that utilizes the 3D skeletal motion data captured using the Kinect depth sensor. The strength of the proposed methods lies in constructing view and scale-invariant feature representations that can effectively capture the underlying spatio-temporal motion patterns of different skeletal joints. A brief overview of the contributions of this thesis is as follows:

1. Two new features are introduced for gait motion representation, namely the joint relative distance (JRD) and joint relative angle (JRA). Both the JRD and JRA features are robust against view and scale variations and can effectively capture the underlying motion patterns of different skeletal joint-pairs (published
2. A new method to evaluate the relevance of a particular joint-pair in gait motion representation is presented, which is based on the flatness measure of the corresponding JRD or JRA sequence. Only the most relevant JRD and JRA sequences are utilized in the final gait signature representation to ensure low computational cost as well as high feature distinctiveness (preliminary concepts were published in [14]).

3. A new dynamic time warping (DTW)-based classifier is introduced for gait classification. The proposed classifier can effectively handle the differences in walking speed, thus eliminating the need of extra pre-processing steps such as resampling (published in [7, 14]).

4. For human action recognition, the joint relative angle (JRA) features are extended to construct a new 2D joint-triplet motion image, which effectively maps the spatio-temporal characteristics of different actions into a compact spatial domain without discarding any sensor data (published in [31]).

5. A local binary pattern (LBP)-based texture descriptor is presented that highlights micro-level texture differences in the joint-triplet motion images constructed for different action classes (published in [31]).

6. To facilitate computational efficiency and high class separation, the Fisher linear discriminant (FLD) method is evaluated. The FLD method projects a given feature space to a low-dimensional sub-space, effectively increasing the between-class discrimination and decreasing the within-class distance.

The effectiveness of the gait and action recognition methodologies presented in this thesis is evaluated using several publicly available Kinect gait and action databases. For evaluating the gait recognition performance, the UPCV gait dataset [1] is used. The UPCV dataset comprises gait video sequences of 30 different subjects captured
using Kinect. For evaluating action recognition performance, two publicly available datasets are used, namely the UTKinect-Action Dataset [2] and the Kinect Activity Recognition Dataset (KARD) [3]. The UTKinect-Action Dataset comprises 10 types of human actions in indoor settings captured using a Kinect sensor. On the other hand, the KARD dataset comprises a total of 18 different activities, which include both gestures and actions performed in an office setting.

The findings and results presented in this thesis can potentially be applied in designing more capable real-time computer vision systems, which can recognize persons from their gait and even track their activities. Augmenting these capabilities with existing security and surveillance systems can potentially enable the system to detect anomaly and trigger alarm. Healthcare can also benefit from such computer vision systems that can be deployed in senior home monitoring to ensure well-being of the residents by tracking their daily activities, or detecting emergency events like fall. In addition, gait and motion analysis can be utilized to assess the movement patterns of sports athletes, with a focus on improving their techniques. The motion features presented in this thesis can potentially be adapted to assess the motion characteristics to improve athletic performance in terms of efficiency. Lastly, the proposed methodologies can also be extended to design intelligent meeting room systems with unobtrusive user authentication and access resource management capabilities, as discussed in my earlier works [11, 30].

1.5 Thesis Outline

The remainder of this thesis is organized as follows. In chapter 2, I first present an overview of the biometric gait recognition and human action recognition approaches. I then explore the related works in both areas and discuss limitations of some of the existing methods. In chapter 3, I present a detailed description of the proposed bio-
metric gait recognition method, which comprises feature extraction, effective feature selection, and the proposed classification and fusion methods. Chapter 4 focuses on the proposed human action recognition method, extending some of the ideas presented in chapter 3. Chapter 5 presents the detailed experimental results conducted to validate the effectiveness of the methodologies proposed in chapter 3 and 4, including dataset description, experimental setup, and comparison with some existing works. Lastly, in chapter 6, I present thesis conclusion, with discussion on the limitations of the proposed methods and possible future works.
Chapter 2

LITERATURE REVIEW

The first section of this chapter gives a brief overview of the Microsoft Kinect sensor with a discussion on some of the precursory works related to range sensing. Next, I present a generic classification of different types of existing gait and action recognition approaches and provide justifications for selecting vision-based marker-less approaches as the primary focus of my thesis. In the subsequent sections, I explore existing approaches to human gait and action recognition including some of the most recent Kinect-based methods. I also discuss the limitations of some of these methods.

2.1 Background

From the advent of computer vision in the early 1960s, researchers are investigating how computers can acquire a realistic interpretation of the complex three-dimensional world around us. One of the main objectives is to enable computers to make sense of a complex environment with the help of sensory input processing, which in turn, has the potential to be utilized in situation-aware intelligent surveillance and access control systems in a natural and unobtrusive manner. Early computer vision methods were mostly based on 2D intensity images, where mathematical techniques were used to model 3D shapes and structures. However, recovering 3D information from 2D intensity images is a challenging task, since projecting a 3D scene to a 2D space incurs significant loss of data. In fact, it is mathematically impossible to construct a 3D representation of an object given only a single 2D intensity image \[32\]. Hence, researchers are actively seeking new sensor technologies in order to recover 3D shapes directly from the sensor, leading to more robust representations and interpretations.
of 3D environments.

Early range sensors introduced in the 1980s were typically based on sonar, infrared, and laser range finders [33]. One of the precursory works on laser range sensors presented by Gil et al. [34] addressed the problem of combining range and intensity data obtained from the same scene by extracting edge maps from both representations and reducing the problem to combining the two edge maps. Magee and Aggarwal [35] presented an extensive review on 3D object description and recognition based on both intensity and laser-based range imagery. They argued that combining the advantages of these two modalities could potentially lead to more robust and computationally inexpensive interpretation and recognition of 3D objects and structures. This argument was further supported by the study presented in [36], where intensity information was utilized to reduce the time required for range sensing. Instead of finding the range for every point in a scene, the authors used intensity image to guide selective range sensing by first detecting potential points of interest. Another work by Vemuri et al. [37] utilized intrinsic surface properties extracted from range data to construct 3D surface and object representations. However, while the range sensors opened the door to a new class of computer vision techniques, early sensors suffered from several limitations. For example, sonar sensors are susceptible to noise caused by echo and reflections. On the other hand, early infrared and laser range finders were expensive and could only estimate the range of a single point in a scene. In addition, typically these sensors were not suitable for capturing human motion data [32].

Being the very first sensor in the category of low-cost consumer-level depth sensing, the recent release of the Microsoft Kinect has potentially opened the door to a new generations of computer vision and biometric security applications [7]. It was originally introduced for the Xbox 360 gaming system as an add-on device that can
detect physical movements or voice commands of the user and thus enable the user to play games without any physical controller [11]. Kinect is made up of an array of sensors, which include i) a color camera, ii) a depth sensor, and iii) a multi-array microphone setup. The depth sensor comprises a monochrome CMOS camera and an infrared (IR) emitter. Using these two components, Kinect can build 3D maps of objects by emitting human eye-invisible IR and then analyzing the light and shadow of the image captured by the CMOS camera. The multi-array microphone has an ambient noise cancellation feature and can also be used to detect the source location of voice. In addition, Kinect can construct a 3D virtual skeleton of a human body using the depth information [38]. With all these capabilities along with its small compact size, the Microsoft Kinect has attracted a significant attention from the computer vision, biometrics, and robotics research community, leading to its application in home monitoring [39], face and facial expression analysis [40], 3D object modeling [41], indoor navigation and mapping [42], healthcare and rehabilitation [43], surveillance [44], etc. In addition, some of the recent works on pose estimation [45], human body modeling [19, 20], motion retrieval [46], and activity recognition [47] have utilized the depth information and computationally inexpensive 3D skeletons obtained from Kinect.

2.2 Overview of Gait and Action Recognition Methods

Human gait and action recognition techniques can be divided into two main categories: i) sensor-based approach and ii) vision-based approach. This section presents a brief overview of these methods.

i) **Sensor-Based Methods:** Sensor-based techniques typically exploit wearable motion sensors attached to different joints or body parts of the subject to measure various characteristics of gait and actions performed by the subject [48, 49]. Some
commonly used wearable sensors include accelerometer, gyroscope, magnetoresistive sensors, flexible goniometer, electromagnetic tracking system (ETS), electromyography (EMG) sensor, etc. [48]. Force sensors and pressure plates that can measure foot pressure have also been successfully applied in human gait analysis [50]. However, although sensor-based techniques can acquire reliable gait and motion-related data, applications are limited to diagnosis of medical conditions and rehabilitation research, typically conducted under controlled laboratory environment [51].

ii) **Vision-Based Methods:** While sensor-based gait and action recognition techniques exploit motion and pressure sensors, vision or image-based systems utilize videos of gait or actions recorded using a single or multi-camera setup in indoor or outdoor environment. The recorded video data is then processed to extract salient characteristics related to human gait or action-specific motion.
This category of gait and action recognition systems can further be subdivided into marker-based and marker-less analysis. In marker-based approach, active or passive markers are attached to body parts of the subject, which facilitates extraction of accurate joint motion data from the video without extensive video processing. On the other hand, in marker-less gait and action analysis, videos are recorded with normal clothing with no marker attached. Different computer vision and image processing techniques are then applied on the recorded videos to extract human silhouette and motion data.

In my thesis, I focus on vision-based marker-less gait and action recognition methods. This choice was motivated by the social acceptability of vision-based approaches, which is well-demonstrated by the widespread deployment and general acceptance of video surveillance systems in public places like airports, banks, bus and train stations, etc. On the other hand, sensor-based approaches are typically difficult to accommodate in many real-world scenarios.
2.3 Vision-Based Gait Recognition Systems

A vision-based gait recognition system involves multistage processing of video data acquired from a single or multi-camera setup. Figure 2.3 shows an overview of a generic video camera-based gait recognition system. As shown in this figure, the first step involves pre-processing the raw sequence images obtained from the camera to isolate human silhouette or motion. Typically, this is achieved by modeling the background scene using a set of sequence images and highlighting the differences in the current image with respect to the background. To facilitate robust background modeling, traditional gait data capturing is performed with a fixed camera position in an environment where the background is relatively uniform [52]. This approach of foreground segmentation is often followed by noise reduction methods to reduce noise and distortion caused by sub-optimal threshold selection [53]. Regular human walking is considered to be a cyclic motion, which repeats in a relatively stable frequency [54]. Hence, the next step involves detecting gait cycles since features extracted from a single gait cycle can represent the complete gait pattern. Once the gait cycle is detected, the corresponding silhouette or motion data is passed to a feature extraction module which extracts the underlying characteristics of individual gait.

Success of a gait recognition system critically depends on the discriminating ability of the extracted feature representation. Hence, it is important to remove redundant and noisy features and utilize only the most informative and discriminating features to construct the final gait signature representation. Lastly, the constructed feature representation is passed to a recognition or verification module that uses machine learning techniques to match unknown gait samples with the training gait models stored in a database.

Depending on the type of features being used, vision-based gait recognition methods found in literature can be divided into two categories: i) model-based approaches
and ii) model-free or appearance-based approaches [55]. In the following subsections, I present brief reviews of the past works in these two categories. In addition, I also present a review of some of the most recent works on Kinect-based gait recognition.

### 2.3.1 Model-Based Gait Recognition

In model-based approaches, movements of different body parts, such as legs, arms, etc. are modeled explicitly based on a set of estimated parameters [54]. The variations of the parametric values are tracked over time, which is then used as the gait signature representation. However, constructing the model, fitting it on the captured gait data, and estimating the parametric values are computationally expensive, which makes model-based gait recognition approaches time-consuming and difficult to accommodate in many real-world applications [54]. BenAbdelkader et al. [56] proposed one of the early model-based gait recognition methods, where two spatio-temporal parameters, namely cadence and stride length were estimated to represent the gait.
biometric. Later, Urtasun and Fua \cite{57} utilized 3D temporal motion model-fitting to synchronized video sequences in their proposed gait recognition method. Individual gait signature was represented based on the estimated motion parameters. A similar approach was adopted by Yam et al. \cite{58}, where gait signatures obtained from walking and running was differentiated by modeling human leg structure and motion. Although their proposed gait recognition method offers view and scale invariance, it depends heavily on the quality of the gait sequences \cite{59}. More recently, Lu et al. \cite{60} introduced layered deformable models (LDM) to represent shapes and dynamics of different human body parts based on 22 different parameters. The parameters were used to capture the size, position, and orientation of the body parts from fronto-parallel gait sequences. While many of the existing model-based gait recognition methods focus on modeling the lower-body parts, the LDM models were applied to construct a full-body representation of gait. Another full-body gait analysis approach presented by Arai and Andrie \cite{61} utilizes morphological operations to obtain skeletal models from extracted silhouettes. Discrete wavelet transformation (DWT) and Haar wavelets were applied on the extracted models to reduce feature dimensionality. However, morphological operation-based skeleton extraction is not invariant against view and scale changes.

2.3.2 Appearance-Based Gait Recognition

While model-based approaches focus on modeling individual body parts and their movements, model-free approaches involve constructing a compact holistic representation of gait motion appearance by utilizing the silhouette sequences extracted from the video \cite{54}. Shutler et al. \cite{62} introduced velocity moment features to represent object and motion in image sequences for gait analysis. In practice, the velocity moments capture the differences between the center of mass of a moving object in successive images. Later, Bobick and Davis \cite{63} proposed the motion energy image
(MEI), which is a temporal template representation of human movement. The MEI representation comprises a static vector image, where each point is a function of the motion attributes of the corresponding spatial location of the point in a sequence image [63]. BenAbdelkader et al. [64] utilized self-similarity plots constructed from pairwise correlation of extracted silhouettes in the image sequences. The obtained plots were then projected into a subspace, namely the EigenGait space using principal component analysis (PCA), which effectively reduces the feature dimensionality. Both the motion energy image (MEI) and the EigenGait based representation of human movement for gait analysis contributed to the development of one of the most popular appearance-based gait recognition methods, namely the gait energy image (GEI) [65]. The GEI is a spatio-temporal representation of all the silhouette motion sequences accumulated in a single energy image [65]. The GEI method utilizes a fusion of principal component analysis (PCA) and multiple discriminant analysis (MDA) [66] to reduce the feature dimensionality, while maintaining a high class separability at the same time.

Some of the more recent model-free gait recognition methods focus on extending GEI to a more robust representation. One example is the frame difference energy image (FDEI) proposed by Chen et al. [67], which handles silhouette incompleteness by utilizing denoising and clustering techniques. Another approach proposed by Li and Chen [68] involves fusing foot energy image (FEI) and head energy image (HEI), which facilitates the construction of a more informative gait signature representation. Nevertheless, although appearance-based approaches present a computationally inexpensive set of gait recognition methodologies, their performance suffer due to scale and view variations in an uncontrolled or changing environment [7].
2.3.3 Gait Recognition Using Kinect

While vision-based biometric gait recognition has been a topic of interest over the past twenty years, the invention of the low-cost Kinect sensor has opened up new opportunities to address the problems related to real-time motion analysis, resulted in a spike in the interest in gait recognition using Kinect. In addition to different data streams that can be obtained from Kinect (RGB, depth, audio), it can also construct a 3D virtual skeleton from human body and track it in real-time \[38\], rendering the traditional video pre-processing tasks (e.g. background modeling, silhouette extraction, etc.) unnecessary. As a result, some of the recent gait recognition methods found in literature utilize the computationally-inexpensive real-time depth sensing and skeleton tracking in order to model the gait signature. One of the precursory work on Kinect-based gait analysis was done in 2012 by Ball et al. \[16\], where the authors utilized features extracted from the lower body skeletal joints to construct the feature descriptor. Unsupervised clustering was used in the experiments to evaluate the effectiveness of the proposed method. Another approach also developed in 2012 by Preis et al. \[69\] utilized 13 biometric features for gait recognition. These features are: height, the length of legs, torso, both lower legs, both thighs, both upper arms, both forearms, step-length, and speed. However, the selected features are mostly related to body structure, often referred as soft biometrics, and do not capture the dynamic motion patterns of the limbs during walking. Gabel et al. \[70\] utilized the differences in the skeletal joint positions between consecutive frames to model human gait. However, the objective was to evaluate gait parameter extraction and therefore, the proposed method was not evaluated for gait-based person recognition.

Apart from using the tracked skeletal joints, some Kinect-based gait recognition methods utilize the depth stream to construct appearance-based gait signature representations which comprise more information than the basic grayscale-based methods.
Sivapalan et al. [71] extended the concept of GEI to 3D and introduced the gait energy volume (GEV) representation for gait recognition. A similar approach was adopted by Hofmann et al. [72], where the depth maps were utilized to extend the GEI to two new depth-based representations, namely the depth-GEI and the depth gradient histogram energy image (DGHEI). In their experiments, the DGHEI-based gait representation achieved better recognition performance, compared to GEI and depth-GEI. Chattopadhyay et al. [73] extended the gait energy volume (GEV) to a more robust pose depth volume (PDV) representation, which utilizes partial volume reconstruction using the RGB and depth streams to obtain more accurate silhouette structure. Depth image-based gait analysis approaches were also introduced for healthcare and home-monitoring in [74] [75].

2.4 Vision-Based Human Action Recognition Systems

A vision-based action recognition system comprises a similar set of components as the generic gait recognition system, as shown in Figure 2.4. However, instead of extracting biometric signature, the focus of an action recognition system is to extract the spatio-temporal nature of the limb movements in order to determine the action class. Augmenting action recognition with gait-based person identification and access control system can potentially enhance the situation-awareness of the system by introducing the ability to detect abnormal behavior and trigger alarm for human inspection. In addition, some of the recent studies have shown that it is also possible to extract soft biometric features from day-to-day activities, which can be incorporated with traditional biometric systems to boost the recognition performance [76] [77].

In this section, I present a brief review of some of the existing vision-based human action recognition systems that utilize model-based and appearance-based features. I also introduce some of the Kinect-based action recognition systems.
2.4.1 Single and Multi-Camera Human Action Recognition

Vision-based action recognition methods found in literature focus on two primary issues: i) finding suitable spatio-temporal features, which include appearance-based global or local features or spatio-temporal interest points, and ii) modeling particular actions based on the found spatio-temporal features [78]. In many cases, the choice of features is dependent on the type of sensor being used. For example, in video-based action recognition, appearance-based motion templates are the most commonly-used features. One of the precursory works on this type of feature representations include the grid-based silhouette appearance features proposed by Yamato et al. [79], which were trained using a hidden Markov model for human action recognition. Bobick and Davis [63] presented a temporal motion template that constitutes two different representations of silhouette motion, namely the motion energy image (MEI) and the motion history image (MHI). Other appearance-based features include optical flow measurements-based motion representation proposed by Efros et al. [80], which
were extracted from spatio-temporal volumes for action recognition at a distance. On the other hand, among the different action recognition methods that utilize spatio-temporal interest points as features, a pioneer work was presented by Laptev [81], where the ideas of Harris and Forstner interest point operators were extended to isolate significant local variations in a space-time domain. The obtained local interest points were successfully applied to detect walking people from video sequences, even under the presence of occlusions and cluttered scenes. Later, Laptev et al. [82] further extended this idea to spatio-temporal bag-of-features for realistic and dynamic action recognition from movie clips. However, although appearance and interest point-based approaches present a computationally inexpensive set of motion representation techniques, their performance suffer due to scale and view variations in uncontrolled or changing environment [7].

While single camera-based action recognition systems focus on extracting appearance-based motion representations and spatio-temporal interest points, multi-camera motion capture (MOCAP) [83] systems construct virtual skeletons and extract 3D joint locations to construct effective model-based motion descriptors. However, such motion capture systems are often dependent on markers and typically expensive to deploy. Natarajan and Nevatia [84] utilized synthetic poses extracted from multiple viewpoint MOCAP data to construct a conditional random field (CRF)-based motion representation. Later, Junejo et al. [85] presented a view-independent action recognition method that utilizes a combination of trajectory-based and image-based self-similarity descriptors and extends the idea of dynamic time warping (DTW) to match two action templates based on self-similarity matrix.

2.4.2 Human Action Recognition Using Kinect

Although the tracked skeletal joints obtained from Kinect are relatively more noisy than the MOCAP data, the computationally-inexpensive nature of the Kinect real-
time skeleton tracking has contributed to some of the recent action recognition methods that utilize the Kinect to construct efficient motion representations. In 2012, Xia et al. [2] utilized histograms of 3D joints (HOJ3D) as compact posture representations for action recognition using Kinect. The extracted posture-based visual words were trained using a hidden Markov model (HMM) on a database of 200 sequences of 10 different activities. Yang and Tian [86] proposed a motion representation based on EigenJoints, which combines static posture and dynamic motion properties from informative frames selected based on accumulated motion energy (AME). A non-parametric Naive-Bayes-Nearest-Neighbor (NBNN) was employed to perform the action classification task. Another approach presented by Theodorakopoulos et al. [87] utilized sparse representation-based classification of human action models, represented using vectors of dissimilarities. The skeletal joint data was first processed to obtain view and scale invariant pose descriptions, which were then classified in a dissimilarity sparse space using k-NN classifier and support vector machine (SVM). Ofli et al. [88] exploited the mean or variance of joint angular motion as a measure to identify relevant or informative skeletal joints to represent specific action classes. Their proposed approach, namely the sequence of the most informative joints (SMIJ), focused on designing an intuitive and easy-to-interpret action motion description, which extracts a combination of joint sequences found to be the most informative or relevant to a particular limb movement involved in performing an action.

2.5 Concluding Remarks

Human gait and action recognition using Kinect is an emerging area of research with majority of the existing works done within the past five years. However, even those few studies were conducted under simplifying assumptions, which include i) using datasets with relatively small number of subjects, ii) considering only the static
shape features instead of utilizing spatio-temporal motion features, iii) being sensitive to changes in view, scale, and walking speed, iv) considering only partial body to extract motion features, etc. Hence, the objective of my thesis is to develop effective and robust methodologies for Kinect-based gait and action recognition that can work in an uncontrolled environment.
Chapter 3

PROPOSED GAIT RECOGNITION METHOD

This chapter presents my proposed Kinect-based 3D gait recognition method that utilizes the skeleton data to construct a robust representation of a human gait signature. The first section provides an overview of the proposed methodology by outlining the components of the system. In the subsequent sections, these components are described in details.

3.1 Overview

The proposed model-based gait recognition system comprises multi-stage processing of the 3D full-body skeleton data obtained from the Kinect sensor. First, complete gait cycles are detected from the skeletal motion sequences obtained from Kinect. Next, I introduce two new features, namely the joint relative distance (JRD) and the joint relative angle (JRA), which are computed over the complete gait cycle. The spatio-temporal JRD and JRA sequences originated from different joint pairs are considered as individual fragments of the holistic gait signature. The main advantage of JRD and JRA-based feature representation is that it is invariant against view and scale changes. As a result, the proposed gait recognition method does not require individuals to walk only in a specific direction (fronto-normal or fronto-parallel) or always maintain a fixed distance from the camera. However, utilizing all joint-pair combinations to construct the feature descriptor is computationally expensive and not all joint-pairs are relevant in gait representation. Selection of relevant JRD and JRA features and classification based on the selected features are challenging since the length of the JRD or JRA sequences might vary for the same person depending
on walking speed. As a result, applying traditional statistical feature selection or classification methods requires resampling of features to obtain equal sized feature vectors extracted from different videos of the same person. However, resampling of time-sequence data involves deleting information or adding pseudo information, which might affect the recognition performance. Hence, I propose a new dynamic time warping (DTW)-based classifier that can process variable length sequences of JRDs and JRAs and compute a dissimilarity measure between the training and the unknown sample without any resampling. In order to assess the relevance of a particular JRD or JRA sequence in gait representation, I propose flatness analysis that evaluates the corresponding joint-pair based on their level of engagement in gait movement. Lastly, the proposed system comprises a fusion module that fuses the JRD and JRA features in score level in order to further boost the recognition performance. To my knowledge, this is one of the first studies that fuses distance and angle-based features for Kinect-based gait recognition. Figure 3.1 illustrates the basic components of my proposed gait recognition system.

3.2 Gait Cycle Detection

Isolating a complete gait cycle is the first task of any gait recognition system, which facilitates extraction of salient gait features. Regular human walking has a cyclic pattern, where different body limbs repeat their movements in a stable frequency [54]. Therefore, features extracted from a complete gait cycle can sufficiently represent human walking pattern without any redundancy. A gait cycle starts with the heal strike of one leg and ends at the next heal strike of the same leg, which includes the following phases in between: foot flat - to - mid-stance - to - heal-off - to - toe-off - to - mid-swing [89]. Therefore, in order to detect a complete gait cycle, the change in the horizontal distance between the two ankles was tracked over time, as shown in Figure
Figure 3.1: Overview of the proposed gait recognition method.

Figure 3.2: Detection of a complete gait cycle based on the Ankle Left and Ankle Right joint distance.
3.2 Intuitively, the distance between the two ankle joints is maximized when the two legs are the farthest apart from each other (a potential heal strike). As a result, three consecutive local maxima potentially corresponds to two consecutive heal strike of the same leg (and one heal strike of the other leg), comprising the complete gait cycle sequence [90]. A moving average filter was applied in order to suppress any noise in the tracked ankle distances over time.

3.3 Gait Feature Representation

The skeleton constructed by the Kinect sensor comprises a hierarchy of 20 skeletal joints, where limbs are represented by connections between joints. The raw data provided by the Kinect is the time series of 3D locations of these joints, which lacks properties like invariance against view and scale changes. In most cases, the raw Kinect data is further processed to obtain a more robust motion representation. I introduce two view and scale invariant gait feature representations: i) joint relative distance (JRD) and ii) joint relative angle (JRA) measures. Although Tang and Leung [91] used a type of JRD (namely, variance JRD) measure for motion retrieval, JRD features have never been used for gait recognition before. On the other hand, most angle-related gait features utilize angles formed by different connected limbs (e.g. angle between the upper and lower leg), where I propose to utilize relative angles formed by different skeletal joints, connected or not. This allows more informative description of the relative movement patterns of different body joints. This section presents a detail discussion on the proposed JRD and JRA features.

Joint relative distance (JRD) between any two skeletal joints $p_1(x_1, y_1, z_1)$ and $p_2(x_2, y_2, z_2)$ can be defined as the Euclidean distance between these two joints in a 3D space [92]:

$$
\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}
$$
\[ \delta(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \] (3.1)

The obtained JRD values are normalized by the height of the skeleton to ensure scale-invariance. JRDs computed over time provide an intuitive representation of the relative movements of the joints involved. This representation is particularly useful in gait recognition, since it can capture the notion of synchronization between different joints. For example, movement of the left ankle during walking is synchronized with the movement of the right ankle. JRDs between these two joints can effectively capture this trait. Previously, Tang and Leung [91] used variance of JRD (VJRD) measures to represent certain human motions in their proposed motion retrieval system. In this thesis, I show that JRDs computed over a complete gait cycle for certain joint pairs can effectively be used to represent the gait pattern. Here, instead of using the variance of JRDs, I treat the JRDs as individual gait signature fragments originated from the corresponding joint pairs.

On the other hand, joint relative angle (JRA) between two joints \( p_1 \) and \( p_2 \) can be defined as the angle formed by \( p_1 \) and \( p_2 \) with respect to a reference point \( r \). Given the coordinates of 3 points \( p_1, p_2, \) and \( r \) in a 3-D space, the angle \( \Theta_{p_1, p_2} \) formed by \( p_1 \rightarrow r \rightarrow p_2 \) using the right hand rule from \( r \) can be calculated as:

\[ \Theta_{p_1, p_2} = \cos^{-1} \left( \frac{\overrightarrow{p_1r} \cdot \overrightarrow{rp_2}}{||\overrightarrow{p_1r}|| ||\overrightarrow{rp_2}||} \right) \] (3.2)

Here, \( \overrightarrow{p_1r} = r - p_1 \), \( \overrightarrow{rp_2} = p_2 - r \), the dot(.) represents dot product between two vectors, and \( ||\overrightarrow{p_1r}|| \) and \( ||\overrightarrow{rp_2}|| \) represent the length of \( \overrightarrow{p_1r} \) and \( \overrightarrow{rp_2} \), respectively. The choice of the reference point can be any other skeletal joint except \( p_1 \) and \( p_2 \). Since the objective here is to encode the relative movement patterns of two joints, the reference point should ideally be a point that remains stationary throughout the entire gait cycle. Hence, the spine joint was selected as the reference point in this
study, since this joint remains almost stationary during walking. Similar to JRDs, changes of JRA values over the full gait cycle also provide an intuitive representation of the relative movements among different joints. The proposed method considers JRDs or JRAs originated from a particular joint-pair as a small fragment of the gait signature. Thus, the full gait signature can be defined as a collection of JRD or JRA values over a complete gait cycle originated from different joint-pair combinations of the human skeleton. For the 20 skeletal joints, there is a total of 190 possible distinct joint-pair combinations, which is a high-dimensional feature space. In addition, not all joint-pair is relevant in gait feature representation. For example, JRDs or JRAs between the spine and the hip center joints does not represent any information related to human gait, since both these joints remain almost stationary when a person walks. Therefore, identifying the skeletal joint-pairs that are relevant to human gait motion is imperative for the proposed gait recognition method.

3.4 Relevant Joint-Pair Selection

Since not all of the skeletal joints engage during human motion, not all JRD or JRA sequences are relevant in gait signature representation. Intuitively, for joint-pairs that have a high relative motion during gait, the joint relative distances or angles computed over the full gait cycle should have frequent spatio-temporal changes. On the other hand, joint-pairs that remain stationary or have very little relative movements during gait should have relatively flat JRD or JRA sequences. Figure 3.3 shows JRD sequences of six different joint-pairs computed over a complete gait cycle for five different persons from the UPCV gait database [1]. It can be observed that for joint-pairs (ankle left, ankle right), (knee left, ankle right), and (knee right, ankle left), the obtained JRD sequences have dominant non-harmonic sinusoidal components, in other words high spatio-temporal changes. This is due to the high relative movements
of these joint-pairs during walking. On the other hand, for joint-pairs like (shoulder left, shoulder right), (hip left, hip right), (shoulder center, hip center), the JRD values are almost constant over the entire gait cycle, which is due to the fact that these joint-pairs have no direct involvement in human walking. Therefore, the shape of a JRD or JRA sequence is an important indicator of the level of engagement of the corresponding joint-pair in human gait. In practice, the more flat the JRD or JRA sequence is, the less relevant it is as a gait feature.

Motivated from the spectral flatness measure (SFM), I propose to utilize a similar flatness measure to quantify the relevance of a JRD or JRA sequence in gait representation. Often called the “tonality coefficient”, the SFM is a measure of the amount of resonant structures (peaks) present in the power spectrum of a signal, as opposed to a flat spectrum [93]. The flatness measure was originally introduced by Gray and Markel [94] for linear prediction of speech analysis, where the SFM was used to quantify the “whiteness” of a signal. In this context, “whiteness” refers to how much noise-like a speech signal is. Since white noise evinces a temporally flat spectrum, selecting an appropriate SFM threshold facilitates differentiating between speech signal and white noise. This concept of spectral flatness has successfully been adopted in different acoustic and speech signal processing problems, such as voice activity recognition [95] and detection of brake squeal [96]. In addition, the MPEG7 framework uses the SFM measure as a low-level feature descriptor for matching and fingerprinting audio signal [93].

The proposed flatness measure of a JRD or JRA sequence can be defined as the ratio of the geometric mean and arithmetic mean of the corresponding sequence, as shown below (the following equation is adopted from [93]):

$$F_i = \frac{\prod_{n=0}^{N-1} JRD_i(n)^{\frac{1}{N}}}{\sum_{n=0}^{N-1} JRD_i(n)}$$

(3.3)
Figure 3.3: JRD sequences for different joint-pairs and persons computed over a complete gait cycle. Here, the x-axis represents frames and y-axis represents normalized JRD values. It can be observed that (a) some JRD sequences have high variations, while (b) the others have a flat pattern.
Here, $F_i$ is the flatness of the JRD (or JRA) sequence obtained for the $i$-th joint-pair. In the context of gait feature representation, the relevance of a joint-pair is inversely proportional to the flatness measure of the corresponding JRD or JRA sequence. A high flatness indicates low relative movements of the corresponding joint-pair during walking, thus yielding a relatively flat JRD or JRA sequence. On the other hand, the flatness measure for joint-pairs that have high relative movements will be low, since the corresponding JRD or JRA sequences will have high spatio-temporal changes, resulted in higher number of resonant structures. Therefore, the proposed method utilizes a threshold on the flatness of JRD or JRA sequences in order to select the most relevant joint-pairs.

3.5 DTW-Based Classifier

Joint relative distances (JRD) or joint relative angles (JRA) for different joint-pairs computed over a full gait cycle essentially represent sequences of time-series data. Alignment of such temporal gait data is a challenging task due to variations of walking speed, which might result in variable length JRD/JRA sequences for the same person. Therefore, applying traditional classifiers in this scenario requires extra pre-processing steps, such as resampling. However, resampling of time-sequence data involves deletion or adding new data, which might affect the recognition performance. One alternative is to utilize non-linear time sequence alignment techniques that can effectively reduce the effect of variable walking speed by warping the time axis. Dynamic time warping (DTW) is a well-known non-linear sequence alignment technique. Originally proposed for speech signal alignment [97], recent DTW applications are mostly verification-oriented, such as off-line signature verification [98]. In this thesis, I propose to utilize DTW to design a classifier for gait recognition that takes a collection of JRD/JRA time series data originated from different joint pairs.
as the parameter and outputs the dissimilarity measure between two given gait samples. Use of DTW in this case allows the alignment of different length JRD/JRA sequences, which enables the classifier to match gait samples without any intermediate resampling stage.

Given the set of all joint relative distances $\text{JRD} = \{\delta_1, \delta_2, ..., \delta_p\}$, where each $\delta_i$ represents JRDs for two particular joints computed over a full gait cycle, I obtain a subset of JRD by selecting relevant joint-pairs using genetic algorithm:

$$\delta = \{\delta_i | i = 1, 2, ..., N \ \text{where} \ \delta_i \in \text{JRD}\} \quad (3.4)$$

Similarly, given the set of all joint relative angles $\text{JRA} = \{\theta_1, \theta_2, ..., \theta_q\}$, where each $\theta_i$ represents JRAs for two particular joints with respect to the reference point computed over a full gait cycle, the selected subset after applying genetic algorithm can be represented as:

$$\theta = \{\theta_i | i = 1, 2, ..., M \ \text{where} \ \theta_i \in \text{JRA}\} \quad (3.5)$$

Let, $\delta_{\text{train}}$ and $\delta_{\text{test}}$ are two JRD sequences from the same joint-pair computed over a complete gait cycle, where the length of $\delta_{\text{train}}$ and $\delta_{\text{test}}$ are represented as $|\delta_{\text{train}}|$ and $|\delta_{\text{test}}|$, respectively.

$$\delta_{\text{train}} = a_1, a_2, a_3, ..., a_{|\delta_{\text{train}}|} \quad (3.6)$$

$$\delta_{\text{test}} = b_1, b_2, b_3, ..., b_{|\delta_{\text{test}}|} \quad (3.7)$$

Here, $a_t$ and $b_t$ are the JRD values of $\delta_{\text{train}}$ and $\delta_{\text{test}}$ at time $t$, respectively. Given these two time series, DTW constructs a warp path $W = w_1, w_2, w_3, ..., w_L$, where $max(|\delta_{\text{train}}|, |\delta_{\text{test}}|) \leq L \leq |\delta_{\text{train}}| + |\delta_{\text{test}}|$. Here, $L$ is the length of the warp path between the two JRD sequences. Each element of the path can be represented as
$w_l = (x, y)$, where $x$ and $y$ are two indices from the $\delta_{\text{train}}$ and $\delta_{\text{test}}$, respectively. There are a number of constraints that DTW must satisfy. Firstly, the warp path must start at $w_1 = (1, 1)$ and end at $w_L = (|\delta_{\text{train}}|, |\delta_{\text{test}}|)$. This, in turn, ensures that every index from the both time series is used in path construction. Secondly, if an index $i$ from $\delta_{\text{train}}$ is matched with an index $j$ from $\delta_{\text{test}}$, it is prohibited to match any index $>i$ with any index $<j$ and vice-versa. This restricts the path from going back in time. Given these restrictions, the optimal warp path can be defined as the minimum distance warp path $\text{dist}_{\text{optimal}}(W)$:

$$\text{dist}_{\text{optimal}}(W) = \min \sum_{l=1}^{L} \{ \text{dist}(w_{li}, w_{lj}) \} \quad (3.8)$$

Here, $w_{li}$ and $w_{lj}$ are two indices from $\delta_{\text{train}}$ and $\delta_{\text{test}}$, respectively and $\text{dist}(w_{li}, w_{lj})$ is the Euclidean distance between $w_{li}$ and $w_{lj}$.

I extend this basic DTW formulation in order to compute the dissimilarity between a training and a testing gait sample, each of which is a collection of JRD/JRA sequences of different joint-pairs. The proposed DTW-based classifier aligns the training and testing JRD/JRA sequences of the same joint-pair with each other and computes a match score between them. Summation of all the match scores obtained for different joint-pair JRD/JRA sequences between the training and testing sample is treated as the final dissimilarity measure. Formally, the final DTW score $\Delta$ for JRD-based gait representation can be defined as:

$$\Delta(\delta, \delta') = \sum_{n=1}^{N} \{ \min \sum_{l=1}^{L} \{ \text{dist}(w_{n,li}, w_{n,lj}) \} \} \quad (3.9)$$

Here, $\delta = \{\delta_1, \delta_2, ... \delta_N\}$ and $\delta' = \{\delta'_1, \delta'_2, ..., \delta'_N\}$ are collections of JRD sequences from $N$ different joint-pairs and $\min \sum_{l=1}^{L} \{ \text{dist}(w_{n,li}, w_{n,lj}) \}$ represents the minimum warp path distance between the $n$-th joint pair JRDs of $\delta$ and $\delta'$. The same formulation can also be used for JRA-based gait representation to compute the final DTW
score $\Delta$: 

$$
\Delta(\theta, \theta') = \sum_{m=1}^{M} \left\{ \min \sum_{l=1}^{L} \{ \text{dist}(w_{m,li}, w_{m,lj}) \} \right\}
$$

(3.10)

Here, $\theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$ and $\theta' = \{\theta'_1, \theta'_2, \ldots, \theta'_M\}$ are collections of JRA sequences from $M$ different joint-pairs and $\min \sum_{l=1}^{L} \{ \text{dist}(w_{m,li}, w_{m,lj}) \}$ represents the minimum warp path distance between the $m$-th joint pair JRA of $\theta$ and $\theta'$.

### 3.6 Score Level Fusion of JRD and JRA

Originally proposed for multi-biometric systems, match score level fusion combines the match scores obtained for different biometric data and is applicable to a wide variety of multi-biometric and multi-feature scenarios [99]. In order to obtain a single match score from either multiple sources of biometric data or multiple feature sets, match score level fusion employs different arithmetic operators, such as addition, subtraction, median, maximum, minimum, etc. on the match scores obtained for individual biometrics or feature sets [100]. In my proposed gait recognition system, individual final DTW match scores obtained for JRD-only and JRA-only recognition are combined using a summation operator and the obtained score is used as the match score for the final decision. Figure 3.4 illustrates the proposed DTW-based classifier and the match score level fusion scheme.

### 3.7 Summary

In this chapter, I present the new Kinect-based gait recognition method, which utilizes the spatio-temporal changes in relative distances and angles among different skeletal joints to represent the gait signature. To evaluate the relevance of a particular JRD or JRA sequence in representing human gait signature, a flatness measure is proposed.
Figure 3.4: Proposed score-level fusion of JRD and JRA features based on DTW classifier.

that reflects the level of engagement of a particular joint-pair in human walking. I also present a new DTW-based classifier that combines individual DTW match scores computed for a collection of JRD or JRA sequences. Here, the use of DTW makes the proposed classifier robust against variable walking speed and thus eliminates any need of extra pre-processing. Finally, a score level fusion of JRD and JRA is proposed to further boost the recognition performance.
Chapter 4

PROPOSED ACTION RECOGNITION

METHOD

This chapter presents the proposed view and scale-invariant human action recognition method that utilizes the Kinect skeletal motion data to construct effective prototypical representations of different actions. The first section of this chapter provides an overview of the proposed methodology by outlining the components of the system. In the subsequent sections, these components are described in details.

4.1 Overview

One fundamental difference between gait and action recognition is the dependence on person-specific motion information. In gait recognition, since the objective is to recognize the person based on his/her walking pattern, JRD and JRA values play important roles in capturing the underlying motion patterns which are dependent on physiological and behavioral trait of the person. On the other hand, in action recognition, the objective is to recognize an action or event which is independent to the person performing the action. Due to the person-independent nature of action recognition problem, I am interested in a more general and high level representation of motion than looking into specific JRD or JRA values (which are typically influenced by person body structure and behavioral trait, just like gait). Hence, for action recognition, I present a new feature representation that captures a more general description of the body limb movements so that person-independent action recognition can be performed.
Figure 4.1: Overview of the proposed action recognition method.
The proposed action recognition method involves several steps. First, the joint relative angle (JRA) features are computed for all skeletal joint-triplets over the complete action sequences. Here, instead of fixing the reference point to always be the hip center, I investigate whether changing this reference point could potentially improve the distinctiveness of the constructed feature representation. Next, individual JRA sequences obtained for different joint-triplets are spatially pooled into a 2D joint-triplet motion image (JTMI), which maps the spatio-temporal characteristics of different actions into a spatial domain without discarding any sensor data. The motivation is to capture the entirety of an action-specific joint movements in a compact prototypical representation, which exhibits different texture responses for different actions. The variations in texture are caused by the differences in the spatio-temporal movements of joints performing different actions. I propose to utilize the local binary pattern (LBP) texture descriptor to highlight micro-level texture differences in the joint-triplet motion images constructed for different action classes. To facilitate effective classification, the resulting features are projected into a discriminant Fisher-space to further increase the inter-class distance and decrease the intra-class variance. Finally, a support vector machine (SVM) is used for the classification task.

The use of a local texture descriptor on the constructed motion images facilitates capturing a more general representation of action-specific joint motions than the specific JRD or JRA sequences, which in turn helps the system to act in a person-independent manner. Figure 4.1 illustrates the different components of the proposed action recognition system.

4.2 Joint-Triplet Motion Image (JTMI)

For action recognition, I propose an extension of the joint relative angle (JRA) feature by considering all possible distinct joint-triplet combinations that can be formed from
the pool of 20 skeletal joints. This in turn, allows to incorporate more information in the constructed feature descriptor, resulting in increased recognition performance.

JRA for a joint-triplet \((j_1, j_2, j_3)\) can be defined as the angle formed by \(j_1\) and \(j_2\) with respect to a reference joint \(j_3\). Given the coordinates of \(j_1\), \(j_2\), and \(j_3\) in a 3-D space, the angle \(\Theta_{j_1,j_2,j_3}\) formed by \(j_1 \rightarrow j_3 \rightarrow j_2\) using the right hand rule from \(j_3\) can be computed as \([30] [7]:\)

\[
\Theta_{j_1,j_2,j_3} = \cos^{-1} \frac{\vec{j_1j_3} \cdot \vec{j_3j_2}}{||\vec{j_1j_3}|| ||\vec{j_3j_2}||}
\]  

(4.1)

Here, \(\vec{j_1j_3} = j_3 - j_1\), \(\vec{j_3j_2} = j_2 - j_3\), the dot (\(\cdot\)) represents dot product between two vectors, and \(||\vec{j_1j_3}||\) and \(||\vec{j_3j_2}||\) represent the length of \(\vec{j_1j_3}\) and \(\vec{j_3j_2}\), respectively.

Temporally, the length of a JRA sequence depends on the frame rate and the time needed to complete a particular action, which might result in variable length JRA sequences, even for the same participant performing the same action. To enable further analysis of the proposed JRA-based action description, the obtained JRA sequences are normalized by time in order to obtain a consistent motion description. Since the proposed method aims to construct a high level representation of action-specific motion, I argue that resampling in this case does not affect the generic distinctiveness of the underlying motion. Hence, I adopt the cubic spline interpolation to resample all the JRA sequences to a fixed length of \(f_r\) frames. Cubic spline interpolation \([101]\) is a computationally inexpensive method for function interpolation, comprising a set of well-conditioned linear equations that facilitates stable and consistent computation, even with large samples.

The proposed method utilizes the spatio-temporal JRA sequences originated from all joint-triplets to construct a holistic spatial motion description, namely the joint-triplet motion image (JTMI). The construction of JTMI allows confronting the problem of spatio-temporal action recognition in a more familiar spatial domain without
discarding any sensor information. First, partial motion images are constructed by spatially concatenating all the JRA sequences for a particular reference joint in a \( f_r \times m \) matrix, where \( f_r \) represents the number of frames in the video sequence (after resampling) and \( m \) represents the number of distinct joint-triplets for a particular reference joint \( i \). An element \( \theta^f_{i,j,jr} \) in this partial motion image represents the JRA value of a particular joint pair \((i, j)\) with respect to the reference joint \( j_r \) in the frame \( f \). The partial motion images obtained for different reference joints are then concatenated to obtain the final joint-triplet motion image (JTMI), which comprises all the joint movements in a compact holistic representation. A pseudocode for the proposed motion image construction is shown in Algorithm 1. Figure 4.2 shows an overview of the process.

**Algorithm 1 Joint-triplet motion image construction**

1: for each reference joint \( i \) do
2:   for each frame \( f \) do
3:     for \( j = i+1 \) to no. of joints do
4:       for \( k = j+1 \) to no. of joints do
5:         Compute the JRA \( \Theta_{j,k,i} \) for the current frame \( f \)
6:       end for
7:     end for
8:   end for
9:   Resample each JRA sequence of \( f \) frames into \( f_r \) frames using cubic spline interpolation
10:  Spatially concatenate all resampled JRA sequences in an \( f_r \times m \) block, where \( m \) is the number of joint-triplet combinations for reference joint \( i \) (= \((n-1) \times (n-2)\), \( n \) = total no. of joints)
11: end for
12: Spatially concatenate all the image blocks obtained for all reference joint \( i \) into the final representation

4.3 Local Binary Pattern (LBP) Texture Descriptor

The local binary pattern (LBP), originally introduced by Ojala et al. [102], is one of the pioneer works on local image texture description. The idea behind the basic
Cubic Spline Interpolation

JRA sequences for different joint-pairs with respect to a reference joint $i$

\[ \Theta_{n-1,n,i}, \Theta_{1,2,i}, \Theta_{1,3,i}, \Theta_{1,4,i}, \Theta_{n-1,n,i}, \Theta_{1,2,i}, \Theta_{1,3,i}, \Theta_{1,4,i}, \Theta_{n-1,n,i} \]

2D Motion Image Block of JRA sequences for a single reference joint. Here, $n$ represents the total number of joints.

Figure 4.2: Construction of the 2D joint-triplet motion image.
LBP operator is to encode the texture information of a small image region into a binary pattern. In order to achieve that, all the neighbors of the local region are compared against the center value. Any neighbor having a gray value greater or equal to the center is encoded as 1 and the rest are encoded as 0. The result is then concatenated to form an 8-bit binary pattern. This process is applied on the entire image using a window mechanism and thus a LBP-encoded image representation is obtained. Formal definition of the LBP operator takes the following form [102]:

\[
LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(i_p - i_c)2^p
\]

(4.2)

\[
s(v) = \begin{cases} 
1, & v \geq 0 \\ 
0, & v < 0 
\end{cases}
\]

(4.3)

Here, \(i_c\) is the gray value of the center pixel \((x_c, y_c)\), \(i_p\) is the gray value of its neighbors, \(P\) is the number of neighbors and \(R\) is the radius of the neighborhood.

Applying the LBP-based texture encoding on a joint-triplet motion representation results in an encoded image representation. Typically, a global histogram computed from this encoded image can be used as a feature descriptor, defined as [103]:

\[
H_{LBP}(i) = \sum_{x=1}^{M} \sum_{y=1}^{N} f(LBP(x, y), i)
\]

(4.4)

\[
f(a, i) = \begin{cases} 
1, & a = i \\ 
0, & otherwise 
\end{cases}
\]

(4.5)

However, a global histogram only represents the occurrence frequency of the local texture patterns and therefore, does not contain any locality information [103, 104, 105]. Hence, I add some notion of locality information of the texture patterns in the feature vector by partitioning the encoded image into equal-sized sub-regions and computing individual local histograms from all those regions. These local histograms
Figure 4.3: (a) The basic LBP encoding scheme, (b) Overview of the LBP feature descriptor construction process.

are then spatially concatenated to obtain the final feature descriptor. Figure 2.4 illustrates the basic LBP feature descriptor construction process.

4.4 Fisher Linear Discriminant Analysis

The proposed method utilizes the Fisher’s linear discriminant (FLD) to further increase the inter-class and decrease the intra-class distances by extracting only the most discriminative features. The FLD method can project a given feature vector (in this case, the LBP feature descriptor) to a low-dimensional feature space and also increase the discrimination between classes as well as decrease the intra-class dissimilarity. For a given training set having \( N \) feature matrices \( \{f_1, f_2, f_3, ..., f_N\} \)
computed for samples from different classes, let us consider that, $c$ denotes the number of classes and each feature matrix is assigned to a particular class $F_k$ ($1 \leq F_k \leq c$). In addition, the mean of the samples from the $i$-th class is represented as $\mu_i$, where $M_i$ represents the number of samples belonging to the class $i$. Based on these notations, we can define the inter-class scatter matrix ($T_B$) and the intra-class scatter matrix ($T_W$) as follows:

$$T_B = \sum_{i=1}^{c} M_i (\mu_i - \mu)(\mu_i - \mu)^T$$  \hspace{1cm} (4.6)

$$T_W = \sum_{i=1}^{c} \sum_{f_k \in F_i} (f_k - \mu_i)(f_k - \mu_i)^T$$  \hspace{1cm} (4.7)

Here, $T_B$ represents the scatteredness of the mean of each class, compared to the mean of all the samples. On the other hand, $T_W$ represents the scatteredness of every feature matrix belonging to a particular class, compared to the corresponding class mean. According to [66], an optimal projection $R_{opt}$ maximizing the ratio of the determinants of $T_B$ and $T_W$ can be represented as a matrix comprising the orthonormal columns for nonsingular $T_W$:

$$R_{opt} = \arg \max \frac{|RT_B R|}{|RT_W R|} = [r_1, r_2, ..., r_p]$$  \hspace{1cm} (4.8)

Here, $\{r_i | i = 1, 2, ..., p\}$ represents the generalized Eigen vectors of $T_B$ and $T_W$. To obtain this, the generalized Eigen value problem is needed to be solved [66]:

$$T_B r_i = \tau_i T_W r_i, \ i = 1, 2, 3, ..., p$$  \hspace{1cm} (4.9)

Here, $\{\tau_i | i = 1, 2, ..., p\}$ represents the $p$-largest generalized Eigen values. Since, there can be at most $c - 1$ nonzero Eigen values, the upper limit of $p$ can at most be $c - 1$. Now, the total scatter matrix $T_s$ can be represented as:
\[ T_s = \sum_{i=1}^{N} (f_i - \mu)(f_i - \mu)^T \]  
(4.10)

However, the intra-class scatter of the projected samples for singular \( T_W \) can be made zero. This problem can be handled by calculating \( R_{opt} \) using the principal component analysis (PCA), as suggested in \cite{66}:

\[
R_{opt}^T = R_{fld}^T R_{PCA}^T
\]
(4.11)

\[
R_{PCA} = \arg \max_R |R^T T_S R|
\]
(4.12)

\[
R_{fld} = \arg \max_R \frac{|R^T R_{PCA}^T T_B R_{PCA} R|}{|R^T R_{PCA}^T T_W R_{PCA} R|}
\]
(4.13)

Here, \( R_{PCA} \) and \( R_{FLD} \) represents the Eigenface and Fisherface projection matrix, respectively. After projecting the LBP feature descriptor using FLD, the most discriminating \( p \) (\( p \leq c - 1 \)) features are selected for the final feature vector. Thus, a more compact and discriminative feature representation is obtained for the proposed action recognition system.

4.5 Summary

In this chapter, I present a new approach to recognize human actions for 3D skeletal motion data captured by the Kinect depth sensor. My contribution includes a concept of extended joint relative angle (JRA) feature, which allows a robust view-invariant motion representation based on the spatio-temporal changes in relative angles among the skeletal joints. Furthermore, a novel spatial holistic description of action-specific motion patterns, namely the 2D joint-triplet motion image, is introduced to better capture intrinsic features of specific actions. The proposed method also utilizes the
local binary pattern (LBP) to isolate prototypical features for different actions. LBP histogram features are then projected into a discriminant Fisher-space, resulting in a compact feature cluster representation to further distinguish individual actions.
Chapter 5

EXPERIMENTAL RESULTS

In this chapter, I present the experimental results for the proposed system. The chapter is divided into two independent sections, which includes the experimental setup, dataset description, and results obtained for the proposed gait recognition and action recognition systems, respectively. In addition, the performance of the proposed methodologies is also compared against some of the existing works.

5.1 Experimental Results for Gait Recognition

This section presents the experimental results for the proposed gait recognition system. A publicly available benchmark dataset, namely the UPCV gait dataset [1] comprising gait sequences of 30 individuals is used to evaluate the proposed system. To quantify the recognition performance, I adopt the cumulative match score (CMS) and the cumulative match characteristic (CMC) curves, computed using a five-fold cross-validation. A detailed description of the UPCV dataset and the selected performance measures is presented in section 5.1.1. Next, I evaluate the performances of the only-JRD and only-JRA based gait recognition. In both cases, I introduce the concept of joint-pair relevance (JPR) matrix, which is utilized to find the most relevant JRD and JRA sequences for gait feature representation. In practice, the JPR matrix is an accumulation of the mean flatness measures of different joint-pair JRD or JRA values computed over the whole dataset. This process is further explained in section 5.1.2 and section 5.1.3, where I also present the results for the only-JRD and only-JRA based gait recognition, respectively. In section 5.1.4, I show the recognition performance for the proposed fusion of JRD and JRA features which is shown to contribute
to a 11.25% increase in the overall rank 1 recognition performance. Lastly, I show that the proposed gait recognition system can attain significantly higher recognition performance than existing Kinect-based gait recognition methods.

5.1.1 Experimental Setup and Dataset Description

The effectiveness of the proposed gait recognition method is evaluated using a publicly available Kinect skeletal gait dataset, namely the UPCV gait dataset. The dataset was created by the University of Patras Computer Vision Group (UPCV) to be used as a benchmark dataset for evaluating biometric gait recognition methods. The UPCV gait dataset comprises gait sequences of a diverse population of 30 participants (15 male and 15 female), who were between the ages of 23 and 55 years during the time of video capturing. For each person, a series of 5 videos was recorded in an indoor environment. Each video comprises skeletal pose sequences of the participant walking in a straight direction captured using a Kinect sensor. The sensor was fixed at 1.70 meters above the ground and was placed at the left side of the participants’ walking path. In this setup, the principal direction of the Kinect sensor formed an angle of approximately 30° with the walking direction. To simulate a natural setup, each participant was instructed to walk at their normal walking speed and the floor was not marked to indicate walking path or provide visual aid. All videos were recorded in three different sessions of the same day, at a frame rate of 30 frames per second (fps). Depending on the walking speed of the participant, the captured gait videos were between 55 and 120 frames.

A five-fold cross-validation is used to evaluate the performance of the proposed method. In a five-fold cross-validation, the entire dataset is divided into five non-overlapping subsets randomly where each subset comprises an equal number of instances for all the classes. The evaluation is conducted in an iterative fashion, where in each iteration the classifier is trained with four subsets and the remaining subset
is used for testing. This process is repeated five times, where each of the five subsets is used for testing once. As a measure of the performance, the cumulative match score (CMS) is computed for each iteration and the final CMS score is obtained by computing average CMS scores of the five iterations. Cumulative match score is a widely used metric for representing the performance of a biometric system, which can formally be defined as [51]:

\[
CMS = \frac{\text{Total Correctly Classified Test Samples}}{\text{Total Number of Test Samples}}
\] (5.1)

Another performance metric that has been used in this thesis is the cumulative match characteristic (CMC) curve, which is an extension of the basic CMS measure. In many real-world biometric systems, a common approach is to generate a list of the top \( n \) candidates for a given probe sample to facilitate further human expert inspection. In this approach, the candidates are ranked based on their match score and the top \( n \)-ranked candidates are provided as the potential match. This in turn, helps to save time since the expert does not need to go through the list of all candidates to verify the identity of a person. Hence, a correct match is considered when a match is found in the top \( n \)-ranked candidates. Thus, a more general definition of CMS for rank \( n \) can be represented as:

\[
CMS_n = \frac{\text{Total Correct Matches in the Top } n \text{ Ranked Candidate List}}{\text{Total Number of Test Samples}}
\] (5.2)
In a CMC curve, the CMS\textsubscript{n} scores obtained for the top \( n \) ranks are plotted in a 2D graph, where the \( x \)-axis represents the ranks and \( y \)-axis represents the CMS score.

5.1.2 Experimental Results for JRD-Based Gait Recognition

This section presents the experimental results for the proposed gait recognition system when only the JRD sequences are used as features. First, I present the concept of the joint-pair relevance (JPR) matrix and explain how this matrix can be utilized to select the most relevant JRD sequences with the help of a threshold \( t \). Next, I present the CMS scores and CMC curves obtained for different thresholds to determine the optimal \( t \) value.

To obtain the person-independent relevance measures for different joint-pair JRD sequences in representing human gait, individual joint-pair relevance (JPR) matrices are constructed for all participants. A JPR matrix, computed for each gait video sequence, is an \( N \times N \) square matrix where a cell \( \{i, j\} \) represents the flatness of the JRD sequence of the \( \{i, j\} \) joint-pair. Here, \( N (= 20) \) represents the total number of joints in the Kinect skeletal model. The individual JPR matrices computed for all video sequences are then utilized to construct a generalized person-independent JPR representation, which is the mean of all the \( 20 \times 20 \) JPR matrices. Figure 5.2 shows a heat map of the generalized JPR matrix computed from the whole UPCV gait dataset that represents the overall relevance of each joint-pair JRD sequence in human gait representation. This matrix is symmetric on the both side of the diagonal, since the JRD values between joints \( \{j1, j2\} \) and \( \{j2, j1\} \) are identical. In addition, the diagonal cells where \( i = j \) are ignored. The generalized JPR heat map yields a comprehensive representation of the relevance of a particular joint-pair JRD sequence in gait feature description, where a large value corresponds to low relevance and a small value corresponds to high relevance. Thus, it facilitates selection of relevant JRD sequences based on a threshold \( t \), where the normalized mean flatness measures
Normalized mean flatness of JRD sequences (computed over the whole UPCV gait dataset)


Figure 5.2: Heat map of the generalized joint-pair relevance matrix computed from the whole UPCV gait database. Here, each cell \((i, j)\) represents the normalized mean flatness of the JRD sequences of joint-pair \(\{i, j\}\).

of the selected joint-pair JRD sequences are less than \(t\).

The performance of the proposed JRD-based gait recognition can be influenced by the selection of the threshold \(t\) value. Hence, finding the optimal \(t\) value that ensures the inclusion of the most relevant joint-pair JRD sequences in the gait feature representation is an important task. In this thesis, the performance of the JRD-based gait recognition is evaluated for different \(t\) values to find the optimal threshold. For a real-time system, this process can be carried out automatically offline before deploying the system and thus the threshold selection process does not contribute to the computational complexity of the proposed gait recognition system. Figure
Figure 5.3: Number of relevant joint-pairs for different threshold ($t$) values. For a given $t$, only the JRD sequences that has the normalized mean flatness equal or less than $t$ is selected for the final gait feature representation.

Table 5.3 shows the number of selected joint-pair combinations for different $t$ values. The selected joint-pairs are shown in Figure 5.4 using JPR matrix representations. In this representation, any cell with a value of 0 corresponds to joint-pairs which are not considered in the final gait feature representation.

The rank 1 cumulative match scores obtained for different threshold values are shown in Figure 5.5. It can be observed that the best recognition performance of 79.33% is obtained for both $t = 0.6$ and $t = 0.7$. However, the number of joint-pair JRD sequences selected for $t = 0.6$ is smaller than $t = 0.7$. This indicates that the gait feature representation obtained for $t = 0.7$ might contain redundant information since the same recognition performance can be attained by a smaller subset of joint-pairs. Increasing the threshold value further results in a decrease in the recognition performance. To further inspect the recognition performance of the proposed JRD-based gait recognition method, I select four $t$ values for which the top four CMS scores are obtained and compute the cumulative match characteristic.
Figure 5.4: Selected joint-pairs based on the normalized mean flatness measure (obtained from the generalized JPR matrix) for different threshold values. Here, the cells with value 0 (dark blue) corresponds to the joint-pairs which are not used in the final feature representation.
Figure 5.5: Rank 1 recognition performance of the proposed JRD-based gait recognition for different threshold values.

(CMC) curve for them. Figure 5.6 shows the CMC curve for the selected threshold values of 0.5, 0.6, 0.7, and 0.8, which provides a more comprehensive representation of the performance of the proposed JRD-based gait recognition. It can be observed that, both $t = 0.6$ and $t = 0.7$ achieve the highest recognition performance for all ranks. The CMS scores obtained for $t = 0.6$ and $t = 0.7$ are almost identical except $t = 0.6$ having slightly lower recognition performance than $t = 0.7$ for rank 3. For both these threshold values, a CMS score of 90% is achieved within rank 5, while selecting any other threshold results in a CMS score of at most 89% for rank 5.

5.1.3 Experimental Results for JRA-Based Gait Recognition

This section presents the experimental results for the proposed gait recognition system when only the JRA sequences are used as features. First, I present the concept of the joint-pair relevance (JPR) matrix for JRA sequences, which can be utilized to select the most relevant JRA sequences with the help of a threshold $t$. Next, I present the CMS scores and CMC curves obtained for different thresholds to determine the
Figure 5.6: CMC curve for JRD-based gait recognition using different threshold (t) values.

optimal t value.

Similar to JRD-based gait analysis, a generalized joint-pair relevance matrix is constructed based on the normalized mean flatness measures of all joint-pair JRA sequences, computed over the whole dataset. First, individual joint-pair relevance (JPR) matrices are constructed for all participants. A JPR matrix, computed for each gait video sequence, is an $N \times N$ square matrix where a cell $\{i, j\}$ represents the flatness of the JRA sequence of the $\{i, j\}$ joint-pair. Here, $N (= 20)$ represents the total number of joints in the Kinect skeletal model. The individual JPR matrices computed for all video sequences are then utilized to construct a generalized person-independent JPR representation, which is the mean of all the $20 \times 20$ JPR matrices. Figure 5.7 shows a heat map of the generalized JPR matrix computed from the whole UPCV gait dataset that represents the overall relevance of each joint-pair JRA sequence in human gait representation. Here, the diagonal cells (where $i = j$), the 11th row and the 11th column are ignored to ensure that each JRA is formed by three distinct joints. The generalized JPR heat map yields a comprehensive representation
Figure 5.7: Heat map of the generalized joint-pair relevance matrix computed from the whole UPCV gait database. Here, each cell \((i,j)\) represents the normalized mean flatness of the JRA sequences of joint-pair \(\{i, j\}\).
of the relevance of a particular joint-pair JRA sequence in gait feature description, where a large value corresponds to low relevance and a small value corresponds to high relevance. Thus, it facilitates selection of relevant JRA sequences based on a threshold $t$, where the normalized mean flatness measures of the selected joint-pair JRA sequences are less than $t$.

The performance of the proposed JRA-based gait recognition is evaluated for different $t$ values to find an optimal threshold. As mentioned in section 5.1.2, for a real-time system, this process can be carried out automatically offline before deploying the system and thus the threshold selection process does not contribute to the computational complexity of the proposed gait recognition system. Figure 5.8 shows the number of selected joint-pair combinations for different $t$ values and Figure 5.9 shows the selected joint-pairs using JPR matrix representations. In this representation, any cell with a value of 0 corresponds to joint-pairs which are not considered in the final gait feature representation. It can be observed that the obtained JRA
Figure 5.9: Selected joint-pairs based on the normalized mean flatness measure of JRA sequences (obtained from the generalized JPR matrix) for different threshold values. Here, the cells with value 0 (dark blue) corresponds to the joint-pairs which are not used in the final feature representation.
sequences demonstrate more temporal changes than JRD sequences, which results in a higher number of JRA sequences with less flatness. For example, while no JRD sequence was selected for \( t = 0.2 \) (as shown in Figure 5.3), more than 50\% of JRA sequences are selected for the same threshold.

The rank 1 cumulative match scores obtained for different threshold values are shown in Figure 5.10. It can be observed that the proposed JRA-based gait recognition can achieve similar recognition performance as JRD-based approach with considerably less number of features. While the JRD-based approach achieves the highest CMS score of 79.33\% with 152 joint-pair JRD sequences, the JRA-based approach can attain a score of 80.67\% with only 71 JRA sequences. This indicates that joint relative angle features are more discriminative in nature than the joint relative distance features. It can also be seen that Increasing the number of JRA sequences by selecting a larger threshold results in gradual decrease in recognition performance. To further investigate the recognition performance of the proposed JRA-based gait recognition, I select four threshold values for which the top four CMS scores are obtained and compute the CMC curve for them. Figure 5.11 shows the CMC curve for the selected threshold values of 0.1, 0.2, 0.3, and 0.4, which provides a more comprehensive representation of the performance of the proposed JRA-based gait recognition. It can be observed that while \( t = 0.1 \) achieves the highest CMS scores for ranks 1 to 3, the best recognition performances for ranks 4 and 5 are obtained when \( t = 0.2 \). Selecting any threshold value other than 0.1 and 0.2 results in a significant decrease in recognition performance.

5.1.4 Fusion of JRD and JRA-Based Gait Recognition

In this section, I present the experimental results for the proposed gait recognition system that utilizes a score level fusion of JRD and JRA features to boost the overall recognition performance. The most relevant JRD and JRA sequences found in section
Figure 5.10: Rank 1 recognition performance of the proposed JRA-based gait recognition for different threshold values.

Figure 5.11: CMC curve for JRA-based gait recognition using different threshold \( (t) \) values.
5.1.2 and 5.1.3, respectively are used in the proposed fusion.

As stated in section 3.6, the proposed gait recognition method utilizes a fusion of JRD and JRA to further boost the overall recognition performance. This performance is also compared against two existing Kinect-based gait recognition methods which utilize the Kinect virtual skeleton to construct gait feature representations. One of the methods was introduced by Ball et al. [16] in 2012 which utilizes the mean, standard deviation, and maximum values of different angles formed by the left and right leg skeletal joints. The other method was introduced by Preis et al. [69] in 2012, where a combination of 14 different features was proposed for gait recognition. Details of the selected methods can be found in [16, 69] and are also reviewed in section 2.3.3. To evaluate these two gait feature representations, I implemented them using MATLAB. A k-nearest neighbor (k-NN) classifier is used for the classification task. Furthermore, to justify the choice of dynamic time warping (DTW) for classifier design, I also evaluate the performance of the proposed method using an Euclidean distance-based classifier instead of DTW, where all gait cycles are normalized to a fixed 40 frames length. Figure 5.12 shows the CMC curves for all the selected methods. It can be observed that fusing JRD and JRA significantly increases the overall recognition performance than only-JRD or only-JRA based methods and outperforms the other two existing approaches. While using only JRD or only JRA sequences for gait feature representation attains a CMS score of around 80% for rank 1, fusing JRD and JRA results in a CMS score of 89.33% for rank 1. The increase in performance is observed for all ranks, with a CMS score of 94% being achieved for rank 5, which is a very high recognition performance for a gait recognition system. In addition, resampling JRD and JRA sequences to a fixed length results in a significant reduction in the overall recognition performance. This result further support the choice of utilizing dynamic time warping (DTW) instead of resampling. Since the proposed
gait recognition exploits person-specific JRD and JRA sequences which are influenced by both behavioral motion and body structure, resampling these sequences often affects the person-dependent features by introducing random noise, adding pseudo-information, or discarding some JRD/JRA values.

From the experimental results, it can be said that Kinect gait recognition based on the DTW classifier and fusion of JRD and JRA sequences is more robust and achieves significantly higher recognition performance than state-of-the-art Kinect-based gait recognition methods. The superiority of the proposed method is due to the novel methodology, which encompasses utilization of fusion with view and scale invariant relative distance and angle features and non-linear alignment of variable length feature sequences using the DTW-based classifier. In addition, the run-time complexity of the proposed DTW-based classifier is close to $O(n^2)$, where $n$ represents the length of the JRD or JRA sequences. Coupled with the proposed relevant joint-pair based feature selection, the DTW-based classifier offers low computational cost, which is
5.2 Experimental Results for Action Recognition

This section presents the experimental results for the proposed action recognition system. Two publicly available benchmark datasets, namely the UTKinect-Action Dataset [2] and the Kinect Activity Recognition Dataset (KARD) [3] are used to evaluate the proposed system. The UTKinect-Action dataset comprises a total of 200 video sequences of 10 different actions performed by 10 individuals, while the KARD dataset comprises a total of 540 video sequences of 18 different actions and gestures performed by 10 individuals. To quantify the recognition performance, I adopt the cumulative match score (CMS) measure, computed using a five-fold cross-validation. A detailed description of these two datasets and the selected performance measure is presented in section 5.2.1. Next, I present the recognition performances of the proposed method for the UTKinect and the KARD datasets in section 5.2.2 and section 5.2.3, respectively. In addition, I also show how the selection of the number of sub-regions of the LBP encoded JTMI image and the selection of resampling rate can impact the recognition performance. To obtain the best performance, I evaluate different parameter values and find the optimal ones. Lastly, the performances of the proposed system are also compared against two well-known action recognition methods. In our experiments, the proposed system achieves very high recognition rates for both datasets, outperforming the state-of-the-art methods by an average of 3.1%.

5.2.1 Experimental Setup and Dataset Description

To evaluate the effectiveness of the proposed method, two publicly available Kinect action databases are used, namely the UTKinect-Action Dataset [2] and the Kinect
Activity Recognition Dataset (KARD) [3]. The UTKinect-Action Dataset comprises 10 types of human actions in indoor settings captured using a Kinect sensor at 30 frames per second (FPS). These actions include: walk, sit down, stand up, pick up, carry, throw, push, pull, wave, and clap hands. Each of these actions was performed 2 times by 10 individuals, which resulted in a total of 20 samples for each action class. The action sequences were recorded from different views (front, back, and right) in order to simulate real-world environment. The resolution of the depth frames was set to $320 \times 240$. Figure 5.13 shows sample RGB and depth frames for different actions from the UTKinect dataset.

The Kinect Activity Recognition Dataset (KARD) dataset comprises 18 different
Figure 5.14: Sample RGB and depth frames from the KARD dataset for different activities [3].
activities, which include both gestures and actions performed in an office setting. The actions include catch cap, toss paper, take umbrella, walk, phone call, drink, sit down, and stand up. On the other hand, the gestures include horizontal arm wave, high arm wave, two hand wave, high throw, draw x, draw tick, forward kick, side kick, bend, and hand clap. All of these activities were performed 3 times by 10 individuals, which resulted in a total of 30 samples for each activity class. The videos were recorded at 30 FPS with a resolution of 640 × 480. Figure 5.14 shows sample RGB and depth frames for different activities from the KARD dataset.

To perform the classification task, a support vector machine (SVM) with a radial basis function (RBF) kernel was used. A five-fold cross-validation was used to evaluate the performance of the proposed method. In a five-fold cross-validation, the entire dataset is divided into five non-overlapping subsets randomly where each subset comprises an equal number of instances for all action categories. The evaluation is conducted in an iterative fashion, where in each iteration the SVM is trained with four subsets and the remaining subset is used for testing. This process is repeated five times, where each of the five subsets is used for testing for once. To evaluate the performance, the cumulative match score (CMS) is computed for each iteration and the final CMS score is obtained by computing the average CMS scores of the five iterations.

5.2.2 Experimental Results for the UTKinect Action Dataset

In this section, I present the experimental results for the UTKinect Action Dataset. Since the performance of the proposed system can be influenced by the selection of the number of sub-regions of the LBP encoded JTMI image and the resampling rate for the JRA sequences, the optimal values for these two parameters are determined first through experimental analysis. The optimal parameter settings are then utilized to evaluate the performance of the proposed system with respect to existing state-of-
The performance of the proposed action recognition method depends on the selection of the number of sub-regions in which the LBP encoded image is partitioned. While partitioning the encoded image into smaller number of sub-regions results in low-dimensional feature space, such representations might lack sufficient information related to the locality of the texture micro-patterns. On the other hand, over-partitioning the encoded image might also affect the overall recognition performance, with added effect of high computational cost caused by a larger feature space. Hence, to find an optimal partitioning of the LBP encoded JTMI image, the performance of the proposed method is first evaluated on different number of sub-regions. Figure 5.15 shows the cumulative match scores obtained for different number of sub-regions. Here, all JRA sequences were resampled to a fixed 40 frame length. It can be observed that the best recognition performance is obtained for a $7 \times 7$ partition of the encoded image. Partitioning to a higher number of sub-regions results in a decrease in recognition performance.

Another factor that can potentially influence the recognition performance of the proposed method is the length of the normalized JRA sequences. Hence, I also evaluate the effect of normalization of JRA sequences at different lengths. Figure 5.16 shows the CMS scores obtained by normalizing the JRA sequences to different fixed-lengths ($f_r$). In all cases, the LBP encoded JTMI images were partitioned into $7 \times 7$ sub-regions to obtain the spatially concatenated histogram features. It can be observed that the highest recognition performance is obtained for $f_r = 45$.

Lastly, the performance of the proposed method is also compared against two other existing action recognition methods, namely histogram of 3d joints (HOJ3D) introduced by Xia et al. [2] in 2012 and EigenJoints introduced by Yan and Tiang [86] in 2014. Both of these approaches utilize skeletal joint coordinate data to con-
Figure 5.15: Recognition performance for the UTKinect dataset for different number of sub-regions into which the LBP encoded JTMI image is partitioned. Only the LBP histogram features are used for the classification task.

Figure 5.16: Recognition performance for the UTKinect dataset for different normalization of JRA sequences. Only the LBP histogram features are used for the classification task.
Figure 5.17: Recognition performance (represented as CMS score (%)) of different methods for the UTKinect dataset.
Figure 5.18: Confusion matrix of the proposed method for the UTKinect Dataset. Rows represent true class and columns represent recognition rate (%).

I implemented these methods using MATLAB. For the proposed action recognition method, the resampled frame length was set to 45 and the LBP encoded image was partitioned into $7 \times 7$ sub-regions. In addition, we also evaluate the performance of only JTMI histogram, LBP encoded JTMI histogram, and LBP encoded JTMI histogram and principal component analysis (PCA) [106]-based feature dimension reduction. Figure 5.17 shows the CMS scores obtained for all these methods. From this figure, it can be observed that the proposed LBP encoded JTMI image histogram coupled with the Fisher linear discriminant analysis achieves the highest overall recognition performance. This also proves the effectiveness of the proposed method against view changes, since the dataset contains action sequences captured from different views. In addition, the figure also shows recognition performances for individual action classes. Here, the proposed method achieves superior performances for all action classes except Sit Down, where both HOJ3D and Eigen-Joints achieve higher CMS scores. Figure 5.18 shows the confusion matrix for the proposed method, which provides a more comprehensive representation of the recog-
nition performances and confusions for individual action classes. It can be observed that the majority of the actions are recognized correctly by the proposed method. There are some confusions between the action-pairs Sit Down and Stand Up, Throw and Pull, and Throw and Clap Hands. Some of the confusions are resulted from the natural similarity of the action-pairs, such as sit down and stand up, and Throw and Wave Hands, since these action-pairs involve similar joint movements. Apart from these, 2 instances of Throw were misclassified as Clap Hands, which might be caused from sensor noise. It can observed from Figure 5.17 that for the action class Throw, all the selected methods achieve relatively lower recognition performances than the other action classes, which indicates possible presence of noise in the dataset for this particular action class.

5.2.3 Experimental Results for the KARD Dataset

This section presents the experimental results obtained for the KARD dataset. The same parameter setting which was found to be optimal in section 5.2.2 was also adopted for the KARD dataset.

Figure 5.19 shows the results of the proposed action recognition method for the KARD dataset. Each bar graph represents the CMS scores obtained for a particular action class. In addition, the overall CMS scores for all action classes are also shown in the last graph. Here also, the performance of the proposed method is compared against the histogram of 3D joints (HOJ3D) [2] and EigenJoints [86] action recognition methods. In addition, I also evaluate the performance of only JTMI histogram features, LBP encoded JTMI histogram features, and LBP encoded JTMI histogram features with principal component analysis (PCA)-based dimensionality reduction. It can be observed that the proposed LBP encoded JTMI histogram coupled with Fisher linear discriminant analysis achieves a perfect 100% CMS score, outperforming all other methods. The second-best recognition performance (CMS score of 98.89%)
Figure 5.19: Recognition performance (represented as CMS score (%)) of different methods for the KARD dataset.

was obtained for the same feature set with PCA-based dimensionality reduction. On the other hand, while the HOJ3D achieves a CMS score of 94.26%, the EignenJoints method attains a score of 96.11%. One particular observation is that although the action classes present in the KARD dataset are similar to that of the UTKinect dataset, the KARD dataset comprises a few additional gesture classes, which might be a reason for the relatively lower recognition performances of the HOJ3D and EigenJoints methods.

From the experimental results, it is evident that the proposed joint-triplet motion image constructed using the extended JRA features can effectively be used to represent and recognize actions and gestures. The superiority of the proposed method is
due to the robust skeletal motion encoding based on the spatio-temporal JRA features, which causes distinctive textural variances in the constructed motion images for different action class. Use of LBP feature descriptor helps to isolate these distinctive features from the images, which are further improved by projecting them into a discriminant Fisher-space. Here, applying the Fisher linear discriminant analysis on the LBP histogram features facilitates low computational complexity by projecting the original feature space into a low-dimensional yet discriminative feature sub-space.

5.3 Summary

In this chapter, I present the experimental results conducted to evaluate the effectiveness of the proposed gait and action recognition methodologies. Several publicly available benchmark datasets are used in the experiments. For both gait and action recognition, the proposed method achieves better recognition performances than some of the existing model-based approaches. For gait recognition, it was found that fusion of JRD and JRA can effectively boost the overall recognition performance and attain a CMS score of 89.33%, while other existing methods attain a maximum score of 60%. On the other hand, for action recognition, the proposed method can achieve a CMS score of around 98% to 100%, where the CMS scores for existing state-of-the-art approaches range from 94% to 96.5%.
6.1 Summary of Thesis Contributions

In this thesis, I present new methodologies for gait and action recognition that utilizes the 3D skeletal motion data captured using the Kinect depth sensor. The strength of the proposed methodologies lies in constructing view and scale-invariant feature representations that can effectively capture the underlying spatio-temporal motion patterns of different skeletal joints. A brief overview of the contributions of this thesis is as follows:

1. Two new features are introduced for gait motion representation, namely the joint relative distance (JRD) and joint relative angle (JRA). Both the JRD and JRA features are robust against view and scale variations. Experimental results from section 5.1.2 and 5.1.3 demonstrate that the proposed features can effectively capture the underlying motion patterns of different skeletal joint-pairs, resulted in high recognition performance.

2. A new method to evaluate the relevance of a particular joint-pair in gait motion representation is presented, which is based on the flatness of the corresponding JRD or JRA sequence. The mean flatness for each JRD or JRA sequences are accumulated in a joint-pair relevance (JPR) matrix, which facilitates selection of the most relevant JRD or JRA sequences for gait recognition. Thus, redundant and low entropy features can be discarded, which facilitates low computational cost as well as high feature distinctiveness. Section 5.1.2 and 5.1.3 shows how the proposed JPR matrix can be utilized to select an optimal subset of JRD and
JRA sequences, respectively.

3. A new dynamic time warping (DTW)-based classifier is introduced for gait classification. The proposed classifier can effectively handle the differences in walking speed, thus eliminating the need of extra pre-processing steps such as resampling. Experimental results presented in section 5.1.2, 5.1.3, and 5.1.4 validates the effectiveness of the proposed DTW classifier. In addition, I have also shown that using resampling on JRD or JRA sequences resulted in a significant decrease in the recognition performance, further supporting the choice of DTW (section 5.1.4). Lastly, experimental results presented in 5.1.4 demonstrate that fusing JRD and JRA features can effectively boost the overall recognition performance, resulted in a 11.25% increase in the CMS score.

4. For human action recognition, the joint relative angle (JRA) features are extended to construct a new 2D joint-triplet motion image, which effectively maps the spatio-temporal characteristics of different actions into a compact spatial domain without discarding any sensor data. Experimental results presented in section 5.2.2 and section 5.2.3 demonstrate the effectiveness of the proposed feature representation on two benchmark datasets. In addition, the high recognition performance obtained for the UTKinect dataset (as shown in section 5.2.2) proves the effectiveness of the proposed descriptor against view changes.

5. A local binary pattern (LBP)-based texture descriptor is presented that highlights micro-level texture differences in the joint-triplet motion images constructed for different action classes. As demonstrated in section 5.2.2 and section 5.2.3, directly using the JTMI image histogram instead of utilizing the LBP texture descriptor results in a decrease in the recognition performance, validating the effectiveness of LBP-based texture feature extraction in the proposed system.

6. To facilitate computational efficiency and high class separation, the Fisher lin-
ear discriminant (FLD) method is evaluated. The FLD method projects a given feature space to a low-dimensional sub-space, effectively increasing the between-class discrimination and decreasing the within-class distance. Incorporating the FLD method in the proposed system facilitates constructing discriminative feature descriptors containing only $N-1$ features, where $N$ represents the number of action classes. Experimental results presented in section 5.2.2 and section 5.2.3 demonstrate that the proposed system achieves the best recognition performance when FLD method is applied on the LBP histogram features, resulting in a very high average recognition rate of around 99%.

6.2 Conclusion

Vision-based gait and action recognition are two of the most widely-studied problems in the areas of biometric and computer vision, with potential applications in security and surveillance, human computer interaction, health-care, intelligent systems, etc. The recent popularization of the Kinect sensor has resulted in a spike in the interest in using the Kinect for gait and action recognition. While the depth sensing capability of Kinect is utilized in some of the existing appearance-based model-free gait and action recognition approaches, only a few study utilizes the computationally inexpensive skeletal tracking to model human gait and action. My work in this thesis focuses on designing new methodologies for Kinect-based gait and action recognition that utilize the 3D virtual skeleton model to construct effective and robust feature representations. While both human gait and action recognition involves analyzing the motion patterns of different body limbs, the objectives are different. In biometric gait recognition, we aim to extract person-dependent motion patterns which are unique to each person, typically caused by the influence of human physiology and behavioral traits. On the other hand, the objective in action recognition is to con-
struct a more generic description of motion which is not influenced by individual physiology or behavior. Hence, the person-independent nature of action recognition requires motion descriptors that can suppress person-specific information while yielding a generic representation. This difference in the objectives of these two problems requires fundamentally different feature representations.

For gait recognition, I introduce an effective gait signature representation that utilizes fusion of two novel features, namely the joint relative distance (JRD) and joint relative angle (JRA). Both JRD and JRA features represent the body limb movements by encoding the relative motion of different joint-pairs through the spatio-temporal changes of distance or angle and can effectively capture the underlying person-specific behavioral and physiological traits related to gait. For example, this description of skeletal joint motion can encode the underlying notion of synchronization between different joint-pairs during human walking, thus facilitating a robust representation of gait signature. Relevance of a particular joint-pair in gait motion representation is evaluated based on the flatness of the corresponding JRD or JRA sequence. My argument is that the flatness of a particular JRD or JRA sequence reflects the level of engagement of the corresponding joint pair in human walking. Only the most relevant JRD or JRA sequences are used in the final feature representation. One of the challenges associated with the proposed approach is how to match variable length JRD and JRA sequences which could be a result of variable walking speed of the same person. To address this challenge, I propose a dynamic time warping (DTW)-based classifier that can effectively match two uneven sequences of JRD or JRA values and compute a match score. The proposed DTW-based classifier facilitates preserving the originality of features by removing the need of resampling, which might discard data or introduce noise. The performance of the proposed method is evaluated using a publicly available Kinect gait database. The experimental results show advantage
of the proposed fusion of JRD and JRA features and DTW classifier based gait recognition over some of the existing approaches.

For action recognition, I introduce a robust view-invariant joint motion representation based on the spatio-temporal changes in relative angles among the different skeletal joint-triplets, namely the extended joint relative angle (JRA). A sequence of JRAs obtained for a particular joint-triplet intuitively represents the level of involvement of those joints in performing a specific action. Collection of all joint-triplet JRA sequences is then utilized to construct a spatial holistic description of action-specific motion patterns, namely the 2D joint-triplet motion image. The proposed method exploits a local texture analysis method, the local binary pattern (LBP), to highlight micro-level texture details in the motion images. This process isolates prototypical features for different actions. LBP histogram features are then projected into a discriminant Fisher-space, resulting in more compact and disjoint feature clusters representing individual actions. The performance of the proposed method is evaluated using two publicly available Kinect action databases. Experimental results validate the effectiveness of the proposed joint-triplet motion image and LBP-based action recognition approach, yielding better recognition performance than some existing methods.

For gait recognition, it was found that fusion of JRD and JRA can effectively boost the overall recognition performance and attain a CMS score of 89.33%, while other existing methods struggle to attain a 60% score. On the other hand, for action recognition, the proposed method can achieve a CMS score of around 98% to 100%, where the CMS scores for existing state-of-the-art approaches range from 94% to 96.5%.
6.3 Limitations and Future Works

Gait features are typically sensitive to changes in clothing and carrying conditions, which makes recognition in dynamic environment a challenging task. In addition, a few studies have shown that it is possible to spoof gait biometric by imitating clothing and selecting individuals with similar physiological builds and attributes \[107, 108\]. However, due to the lack of the availability of any Kinect-based gait spoofing dataset, robustness of the proposed methodology under such attacks was not evaluated in this work. One possible alternative is to incorporate other biometric modalities, such as face or voice, in order to increase the robustness of the system. Another promising direction is to incorporate context information with gait features to boost the recognition performance. A recent study \[109\] on fusion of context-metadata with gait recognition has shown significant improvement over gait-only biometric recognition, where user contextual and behavioral patterns are modeled. User information related to the daily routine such as being at a specific location during a specific time (for example, being in some workplace during the morning) and specific conditions related to the location (for example, working on a desk in an office, attending a meeting, etc.) provide valuable context metadata that can be used to improve the biometric recognition performance. Another example of behavioral context can be a collaborative environment where user behavior related to the manner of communication, situational responses, temporal patterns of user activity, preference of spatial location, and interaction among group members and project organizers can be utilized as metadata \[11\].

On the other hand, activity-related soft biometrics-based authentication has attracted much attention in recent years for its potential applicability in ambient and collaborative intelligent environments \[76\]. It has been shown that activities that involve multi-joint movements contain certain behavioral signatures related to indi-
vidual physiology, style, and perceived environment [77], which can be exploited in an unobtrusive authentication system. Some examples of such biometric traits that have been explored in literature include sensing seat-based unobtrusive anthropometric biometric that models the user weight distribution patterns during sitting action [110], prehension movement-based biometric that models activities such as grasping, reaching, and manipulating a particular object [111], etc. In this context, the proposed gait and action recognition methods could further be evaluated to see whether action-specific soft biometric features could be extracted for individuals. This ability can potentially be augmented with traditional gait recognition, resulted in increased robustness and recognition performance. Finally, a method utilization for immersive environments and virtual reality, as well as medical applications and collaborative environments, will be interesting new domains to explore.
Bibliography


[41] M. Jaiswal, J. Xie, and M. Sun, “3d object modeling with a kinect camera,” in *Asia-Pacific Signal and Information Processing Association Annual Summit*


[65] J. Han and B. Bhanu, “Individual recognition using gait energy image,” *IEEE*


[99] M. He, S. Horng, P. Fan, R. Run, R. Chen, J. Lai, M. Khan, and K. Sentosa,
“Performance evaluation of score level fusion in multimodal biometric systems,”

[100] M. Gavrlova and M. Monwar, Multimodal Biometrics and Intelligent Image


[102] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rota-
tion invariant texture classification with local binary patterns,” IEEE Transac-
tions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987,
2002.

expression recognition,” IET Electronics Letters, vol. 48, no. 19, pp. 1203–1204,
2012.

patterns for facial expression recognition,” in IEEE International Conference

[105] F. Ahmed, P. Paul, M. Gavrilova, and R. Alhajj, “Weighted fusion of bit plane-
specific local image descriptors for facial expression recognition,” in IEEE In-
1852–1857.


