Doctoral Dissertation

Multi-View Gait Analysis in Surveillance Scenes by Temporal Geometric Features of Human Body Parts

Thanyamon PATTANAPISONT

Supervisor: Shinobu HASEGAWA

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September, 2024

Abstract

A gait is a walking pattern that can be used to identify a person. Walking involves changing the whole body's joints and initiating postures. Humans have individual walking postures that depend on their velocity, arm swing, foot placement, etc. It can represent personality, identity, and health conditions that affect walking, such as pain, injuries, and neurological diseases. Understanding the human gait improves an analysis system for clinical, psychological, security, and more. Recently, gait analysis has incorporated a vision-based method, using a camera as a tool to access the gait information. Accordingly, it is a non-complicate, flexible, and cost-effective system. However, it suffers from a view-variation issue that reduces the reliability of a vision-based gait analysis, especially for identification tasks.

The surveillance scenario is crucial as it is included in a real-world situation. It can be applied in general for various purposes, such as security purposes. The reliable identification of video surveillance cameras is essential to improving security. We can identify suspicious individuals through their gait when they appear on the surveillance cameras because the gait is difficult to pretend or change, unlike appearance.

This research aims to propose a method for handling identification in a multiple surveillance camera environment using pattern matching based on the distance calculation method and voting. We apply a majority vote to integrate the information from multiple perspectives to overcome the view-variations problem. Notably, it is not a cross-view recognition, as in the previous studies.

Because the surveillance scenario is uncontrollable, markers cannot be attached to the walker's body. This research implements vision-based human pose estimation algorithms to solve this problem. We applied these algorithms to the human joints on sequences and extracted the features. We propose two approaches according to the features. Approach 1 & 2 are a pattern matching based on Dynamic Time Warping (DTW) with time-dependent features (joint angles and time-dependent correlation), and approach 3 is a pattern matching based on Euclidean distance (EU) with a time-independent

correlation feature. We extract the joint angles and correlation as features based on a skeleton landmark from vision-based pose estimation.

This experiment used the CASIA-B dataset to represent the eye-level scenario and the OUMVLP-Pose dataset to represent the surveillance scenario. Furthermore, we adjust parameters by separating features into three parts, i.e., whole, upper, and lower body, to study the impact of different body parts on gait, and remove each joint one by one to study its importance to the gait analysis. Moreover, we separate the number of subjects in the CASIA-B and OUMVLP-Pose datasets into three cases to study the effect of the data amount on the gait analysis.

For approach 1, the whole body feature (excluding the back ankle) is essential for the eye-level scenario and surveillance scenario when using AlphaPose as a pose estimator, but the lower body feature is sufficient for the surveillance scenario when using OpenPose as an estimator. However, the whole body feature is critical for approach 3. Furthermore, approach 1 is the most suitable to apply with gait because it maintains time information and DTW allows time warping. This makes approach 1 better at handling a situation when the same person is walking at a different speed. We found that approach 2 is unable to be employed for identification due to insufficient data variations.

In addition, we determined the significance of each joint and found that the back ankle is a noise (for the eye-level scenario). We can increase the accuracy by removing it from a feature vector. We conducted the experiment by using weighted voting instead of majority voting. The results prove that a majority vote improves the view-variation issue by integrating different perspectives, which is better than a weighted vote.

Compared with the existing studies, our approaches produce a competitive result, especially for the surveillance scenario that is our main focus. Furthermore, the results indicate that pattern matching can perform the identification task on a small database and provide flexibility when changing the database's quantity. It suggests that pattern matching is an alternative method for accessing human gait.

Keywords: multi-view gait analysis, joints feature, distance calculation, pattern matching, voting algorithm

Acknowledgement

I would like to express my deepest appreciation to my supervisors, Professor Kazunori KOTANI, Professor Shinobu HASEGAWA, and Assistant Professor Prarinya SIRITANAWAN, for their great suggestions, important discussions, and support. Their guidance encourages me to keep researching from the beginning to this moment.

I thank all committee members, Associate Professor Kiyoaki SHIRAI, Professor Kokolo IKEDA, and Associate Professor ABE Toru (from TohoKu University) for their constructive comments and suggestions to improve dissertation.

I would also wish to express my sincere gratitude to Associate Professor Toshiaki KONDO, Associate Professor Waree KONGPRAWECHNON, and Dr. Jessada KARNJANA, for their encouragement and helpful feedback.

I am grateful to all of my colleagues at JAIST for their mental and physical supports, inspiration, and motivation. Thank you for constantly sharing excellent dishes and keeping me with you at all moments. My life at JAIST would be difficult without them.

Scholarship and financial supports from JAIST, SIIT, and NSTDA are acknowledged.

Lastly, I desire to express my heartfelt thanks to my parents, relatives, and all of my Thai friends for their kind and invaluable mental support, which has motivated me to remain on track till graduation.

List of publications

1. **Journal Papers**

   - T. Pattanapisont, K. Kotani, P. Siritanawan, T. Kondo, and J. Karnjana. 2024. "Multi-View Gait Analysis by Temporal Geometric Features of Human Body Parts," *Journal of Imaging*, volume 10, issue no. 4, 88. doi: 10.3390/jimaging10040088

2. **International Proceedings**

   - T. Pattanapisont, T. Leelasawassuk, W. Kongprawechnon, T. Kondo and H. Shoichi, "Localization of Generator Inspection Vehicle using Visual Odometry" *2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)* 2019, pp. 400-404, doi: 10.23919/SICE.2019.8859867.
   - T. Pattanapisont, T. Kondo, K. Kotani, J. Karnjana, and W. Kongprawechnon, "A comparative study between temporal Median filter and accumulative difference image for moving object extraction," *International Conference on Instrumentation, Control, Information Technology and System Integration (SICE Annual Conference 2022)*, 2022, pp. 366-369.
   - K. Sirinaksomboon, P. Tantawanich, T. Pattanapisont and T. Kondo, "Moving Object Extraction Based on the Accumulative Differences of Image Intensities and Gradient Orientations," *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, 2023, pp. 516-521, doi: 10.23919/SICE59929.2023.10354111.
   - P. Tantawanich, K. Sirinaksomboon, T. Pattanapisont and T. Kondo, "Moving Object Extraction Based on the Temporal Median Filtering for Image Intensity and Gradient Orientation Information," *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, 2023, pp. 510-515, doi: 10.23919/SICE59929.2023.10354114.

- T. Pattanapisont, P. Siritanawan, K. Kotani, T. Kondo and J. Karnjana, "Gait Image Analysis Based on Human Body Parts Model," *2023 IEEE International Conference on Agents (ICA)*, 2023, pp. 24-27. doi: 10.1109/ICA58824.2023.00014.

3. **Domestic Proceedings**

- T. Pattanapisont, S. Prarinya and K. Kazunori, "Gait Recognition by the Voting Method on Temporal Geometric Features of Human Body Parts," *Image Coding Symposium and the Imaging Media Processing Symposium (PJSM/IMPS 2023)*, Gotemba, 2023.
- T. Pattanapisont, K. Kazunori and S. Prarinya, "Gait image analysis by the voting method on the body parts feature," *307th Research Meeting of the Institute of Image Electronics Engineers in Ishikawa*, Ishikawa, 2024.

Contents

This dissertation was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Sirindhorn International Institute of Technology, Thammasat University.

List of Figures

List of Tables

Chapter 1

Introduction

Walking is a basic mode of transportation for humans. It is a relationship between bones, joints, and muscles that interact with each other, and the nervous and brain systems relate to these activities, and the gait is an individual's walking pattern that involves position changes in the upper and lower body. It refers to the movement of joints and muscles as they change position over time when we take a step. It appears to be a simple behavior that occurs in our daily lives, but the gait or walking pattern provides more insight into individual information than the direction and destination we are heading to.

Our walking pattern, similar to the face, iris, and finger print, can represent a person's personality as shown in Figure 1.1. These unique movements can serve as a representative of our identity, physiology condition, and overall well-being, as suggested by Singh et al. [42].

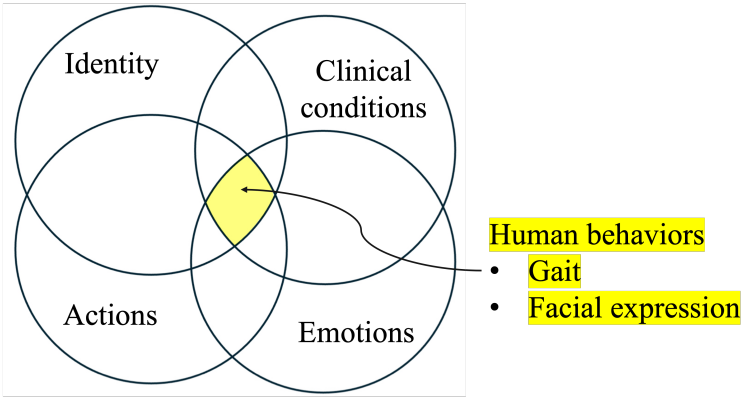


Figure 1.1: Visualization of human personality definition.

The significant changes in walking pattern, such as an irregular arm swing

or body asymmetry, may indicate a medical condition like balance or coordination issues. Furthermore, it can show other diseases, such as neurological disorders. Additionally, our gait reflects our emotions and psychological conditions. Human feelings affect posture and gait differently as many previous studies indicated that our feelings are visible through our walking patterns. This knowledge benefits widespread areas, e.g., security, re-identification, clinical examination, and emotion recognition.

For humans, gait is more complex than a simple mode of transportation. It provides information about an individual beyond their walking style, including their personality, emotions, and health. Understanding the gait should have positive impacts. Figure 1.2 presents the diagram of the gait analysis tasks that digests the gait into various information depending on purposes.

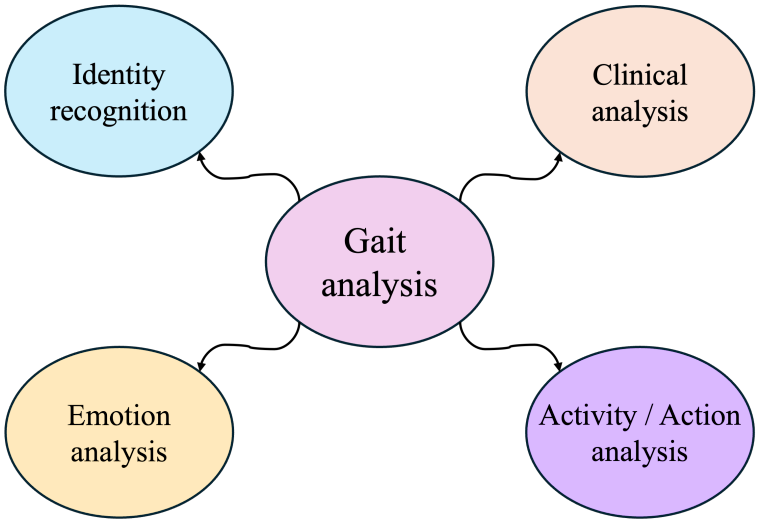


Figure 1.2: Diagram to presents the gait analysis purposes.

1.1 Background of the gait analysis

This section presents the background and previous studies on gait analysis. In clinical analysis, there are various tasks that use gait as the main information, as well as emotions, activity or action, and identity recognition. One of the most familiar studies about gait is related to clinical examination tasks. It uses gait to examine abnormalities such as pain, injuries, and diseases that affect changes in walking patterns. The gait-based emotion analysis uses the walking pattern to identify the walker's feelings, including happiness, sadness,

fear, anger, and neutrality. It is proof that our emotions affect our walking patterns. Activity and action recognition is an analysis that uses a walking pattern, or gait, to predict the target's activity. Additionally, gait-based identity recognition has recently become a widespread and active research topic. It is an analysis that uses gait to identify the walker's identity, which is mostly vision-based. Figure 1.3 summarizes the main purposes of the gait analysis. Mostly, the classification is applied to analyze complex information from gait, especially for medical conditions and emotion analysis, because the classification is mostly related to the neural networks. Meanwhile, pattern matching is more suitable for identity or personal analysis. The following subsections included the categorized previous studies for presenting what has been done in this research field.

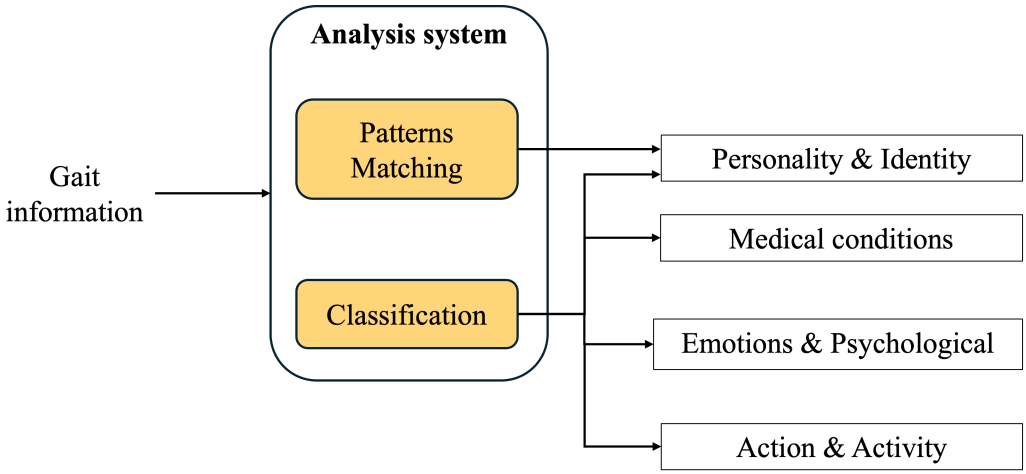


Figure 1.3: Visualization of the analysis system for analyzing the human gait.

1.1.1 Gait in clinical analysis

Gait analysis can identify some neurological diseases, such as Parkinson's disease (PD), which affect a walking pattern. Parkinson's disease (PD) is a progressive disorder that affects the nervous system. The walking pattern presents some of PD's symptoms. The study by S. R. Hundza et al. [26] showed how to use an Inertia Measurement Unit (IMU) to find the first steps of people with Parkinson's disease. They did this by reversing a gyroscope's angular rate to examine their gait cycle. A. P. Rocha et al. [40] employ the Kinect RGB-D camera system as a tool to assess PD by extracting the

skeleton of PD patients. Their goal was to distinguish between PD and non-PD subjects, as well as between two PD states.

In addition to adults and elderly diseases, children can suffer from damage that affects brain development, which is known as cerebral palsy (CP). It typically occurs before birth and affects children's movement and posture. D. Slijepcevic et al. [44] used different machine learning (ML) and deep neural networks (DNN) techniques to classify the walking patterns of children who have CP. They aimed for explainable ML to gain trust in using ML to analyze human gait. They found that the classification from ML approaches is better than DNNs. However, DNNs employed additional features to predict the results.

Moreover, researchers can use gait analysis to assess the risk of falling, thereby preventing potentially serious injuries. There is research using gait analysis to detect a fall state in adults and the elderly, as in the paper from G. Sun and Z. Wang [49]. They suggested using vision-based fall detection by OpenPose to figure out what the human pose is, as well as applying SSD mobilenet object recognition to get rid of OpenPose's mistakes. Then, apply the SVDD classification for fall detection.

Other research related to the gait for clinical analysis has been studied, such as the work from Y. H. Yeh et al. [54] that proposed the method to analyze the frequency domain of the IMU acceleration signal by applying the Discrete Fourier Series to detect the gait cycle time (GCT). They suggested that GCT is in between the heel strike and toe-off sub-phases in the human gait cycle, and their method can detect this information. Meanwhile, the ML model has been employed with gait analysis and is aimed at classifying or identifying abnormalities in the patients. However, the ML model suffers from trust issues due to its black box characteristic, which is unable to explain the reasons for the obtained results. D. Slijepcevic et al. [43] aimed to enhance the visibility of the black box characteristic of neural networks by explainable artificial intelligence (XAI). They selected the layer-wise relevance propagation (LRP) method to obtain the explanation from multi-classification techniques such as the Convolutional Neural Network (CNN), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). The signal they analyzed is based on Ground Reaction Force (GRF), which is a wearable force sensor to measure the force that reacted to the ground. They claimed that XAI can detect the bias of the ML model, which is crucial for gaining the trust of the automatic gait analysis model. Vision-based gait analysis has become an active research topic in clinical analysis. Since it requires less equipment than wearable sensor-based devices and is non-contact with patients, it remains an efficient and reliable analysis outcome. Since vision-based gait analysis mostly relies on a visual pose estimation algorithm,

which will mark the estimated human joints on the captured sequences, S. Jan et al. established a video-based clinical gait analysis workflow that can deploy videos from smart phones or tablets without prior gait analysis expertise [48]. They employed OpenPose pose estimation to estimate 2D human landmarks from the frontal and sagittal planes. Then, they estimated depth, step length, and gait speed by using trigonometric relationships. They found that this proposed method produces the same accurate results as the one from the 3D motion capture system, and it can perform an analysis over different groups of patients, such as adults with normal gait, post-stroke patients, and Parkinson's disease patients. A. Cimorelli et al. proposed a validation study on a video-based clinical gait analysis for prosthesis users [11]. Since the general pose estimation algorithms are employed with general human anatomy, their performance has become poor with prosthesis users. However, they improved it by training a prosthetic-specific joint detector to let it work on prosthesis users, and they claimed that it performed better than MMPose with the COCO dataset. Moreover, they validated the results of gait parameters based on video compared with the data obtained from IMU. Their proposed method produces results close to the values measured by the IMU. Furthermore, H. Chang Soon et al. published a research article on automated gait analysis based on a pose estimation algorithm [22]. They employed MediaPipe pose estimation to detect the key gait events in gait videos, which are heel strike and toe-off. After that, calculate the gait parameters, such as stance time, swing time, step time, and double support time, compared with Vicon moCap, the motion capture system. They found that the gait parameters extracted from the pose estimation algorithm were satisfied when compared to the motion capture system. However, they found that it still produces false detection that affects the missing and incorrect values in the gait parameter calculation.

1.1.2 Gait in emotion analysis

There are many studies that prove the human gait can be used to detect emotions, as per the survey from S. Xu et al. [53]. The study from G. E. Kang et al. [28] looked at how bipolar disorder patients control their balance while walking and sitting to walk. To do this, they used motion data from 16 cameras. Y. Bhatia et al. [6] adopted Long Short-Term Memory (LSTM) and MLP models to recognize four emotions, i.e., happiness, sadness, anger, and neutral, through the gait. They provided joint coordinates based on motion capture as an input to the networks for classifying these four emotions. They claimed that the proposed method performed better and required less inference time than other gait-based emotion recognition

methods. Moreover, N. Jianwattanapaisarn [27] conducted a study to analyze an emotion characteristic by providing 49 subjects to walk in a setting region while watching the emotion-inducing videos on Microsoft Hololens 2 smart glasses. They used OptiTrack motion capture to obtain human gaits and postures, and they extracted features such as the angle between body parts and walking straightness for analysis. A previous study by C. Song et al. [45] proposed a self-supervised gait-based emotion representation (SSAL) to recognize the emotion through the gait. They fed a 3D human skeleton into the input and used Selective Strong Augmentation (SSA) to predict the class of emotion from unlabeled data. They created the SSA, which aimed to improve the model's performance and acquire more resilient features from positive samples. Then, they employed the complementary feature fusion network (CFFN) to extract the features, which are a fusion between structural and representative features. Their proposed method suffered from the unbalanced emotion label in the gait dataset. C. Bisogni et al. [45] also mentioned an unbalanced gait emotion dataset in their paper. They constructed a framework for recognizing emotions based on gait called "Walk-as-you-Feel" (WayF). This approach focused on skeleton sequence analysis and avoided using facial features, which aimed to retain the privacy of walkers. When using an unbalanced dataset, their method incorrectly classifies "sad" emotions. However, they suggested that excluding "neutral" increases accuracy.

1.1.3 Gait in activity analysis

This sub-section presents previous studies about gait and activity recognition. Since gait can recognize walking, it is one of the activities performed by humans. We can use it to categorize related activities, such as running, jumping, and jogging.

The paper from J. Gupta et al. [19] proposed a vision-based activity recognition through gait to identify the performed activity, such as walking, running, jogging, or jumping, by a movement of human legs. They utilized Hu-moments to determine the centroid of the human body. Then, they applied the Mean-shift algorithm to recognize the leg component. Finally, they extracted and classified four activities based on the features extracted from the leg components. H. Chidananda and T. Hanumantha Reddy [20] presented the method to recognize human activity based on foot movement patterns in the gait sequences. They performed human tracking to extract human sequences and find the foot points based on the human body's threshold, on which only the lower part was focused. Then, they determined distance and angle and classified four activities, i.e., walking, running, jogging, and jumping, based on the extracted features. Meanwhile, P. Srihari and J.

Harikiran [47] performed activity prediction based on the human skeleton of thermal images using Siamese networks. They employed PoseNet to identify the human pose and used Siamese networks to determine the similarity between images. They used the similarity score to predict activities.

1.1.4 Gait in identity recognition

The following prior studies suggest that we can identify a person's identity using gait.

The work from M. Alotaibi and A. Mahmood [1] intends to increase gait recognition accuracy by developing eight layers of deep CNN that are less sensitive to variations and occlusions. They employed CASIA-B, a multi-view gait database with various walking conditions, for the experiment. Their proposed method can overcome several issues, but the performance will decrease if the gallery set does not cover a variety of walking conditions. They achieve an average correct classification, rank-1, and rank-5 accuracy of 86.70%, 85.51% and 96.21% on the CASIA-B dataset, respectively. M. Deng and C. Wang focus on proposing gait recognition in different clothing conditions [12]. They employ silhouette gait images to extract the shape of a human and divide it into four sub-regions. Then, select the gait features based on the width of each sub-region and input the gait feature vector into Radial Basis Function (RBF) neural networks. Their proposed method returns the correct classification rate on NM and CL conditions of the CASIA-B dataset as 90% when using NM as a probe set and 93.5% when using the CL condition as a probe set. S. Hou et al. developed the Gait Lateral Network (GLN) to recognize the human gait [23]. It is a deep CNN that can learn discriminative and compact representations from silhouette images. GLN achieves average rank-1 accuracy of 96.88% on NM and 94.04% on BG condition of CASIA-B dataset, respectively. However, the clothing condition affects the slight decrease in rank-1 accuracy to 77.50%. C. Fan et al. [13] claim that different parts of the human body consist of diverse visual appearances and movement patterns during walking. GaitPart was proposed as a way to extract gait features. The goal is to improve the learning of part-level features using a frame-level part feature extractor made up of FConv and get the short-range spatiotemporal expression by using a Temporal Feature Aggregator with a Micromotion Capture Module (MCM). The results from GaitPart achieve average rank-1 accuracy on the CASIA-B dataset as 96.2% on NM, 91.5% on BG, 78.7% on CL conditions, and 88.7% on the OU-MVLP dataset. GaitEdge is a framework described by J. Liang et al. [32] for recognizing human gait. It can make this framework more practical and keep performance from dropping in cross-domain situations by blocking irrelevant

gait information. They designed the module to integrate the trainable edges of the segmented person's shape with the fixed internals of silhouette images based on the mask operation, named Gait Synthesis. GaitEdge achieves the average Rank-1 accuracy on the CASIA-B* dataset (across different views) of 97.9% on NM, 96.1% on BG, and 86.4% on CL conditions.

The early works from R. Liao et al. [34] proposed a model-based gait recognition by extracting 14 body joints of 2D human pose estimation from images and transforming them into 3D poses, called PoseGait. The CNN is implemented to extract the gait features. Moreover, they combine three spatio-temporal features with the body pose to enhance the features and recognition rate. Their proposed method achieves recongition rates on the CASIA-B dataset of 63.78% on NM, 42.52% on BG, and 31.98% on CL conditions. Additionally, they proposed another model-based method for gait recognition with pose estimation and graph convolutional networks, named PoseMapGait [33]. They aimed to preserve the robustness against human shape and the human body cues of the gait features by using a pose estimation map, which claimed to enrich the recognition rate. PoseMapGait achieves the average recognition rate on the CASIA-B dataset as 75.7% on NM, 58.1% on BG, and 41.2% on CL conditions. X. Li et al. [30] mentioned the information loss suffering of 2D poses, unlike 3D poses, which have richer pose information. They present a 3D human mesh model with parametric pose and shape features. In addition, they trained a multi-view to overcome the poor pose estimation in 3D space. They achieve Rank-1 accuracy on the CASIA-B dataset as 60.92% on NM, 42.01% on BG, and 32.81% on CL conditions. This study is not trained for gait recognition directly, but they aim to create the database for multiple related purposes. The research from C. Xu et al. [52] considered the occlusion-aware human mesh model for gait recognition. They mentioned that a partial occlusion of the human body mostly occurred in surveillance scenes. So, they create model-based gait recognition for handling the occluded gait sequences without any prerequisite. They set the SMPL-based human mesh model to an input image directly, extracting the pose and shape features for the recognition task. The most challenging part is when the occluded ratio is huge (around 60%). Their proposed method outperforms the other state-of-the-art methods by 15% of the rank-1 accuracy. K. Han et al. proposed a discontinuous gait image recognition based on the extracted keypoints of the human skeleton [21]. They aim to overcome the situation of discontinuity in the gait images. This study achieves a high recognition rate and is robust to common variations. Mostly, model-based gait analysis aims to increase the recognition rate by implementing machine learning. They achieve the average Rank-1 accuracy on 3 conditions of the CASIA-B dataset as 79.5%.

Previous studies have addressed various variations that make gait analysis unreliable, such as camera perspective, clothing, illumination, occlusion, and carrying. These variations are significant challenges for analyzing the gait. Furthermore, it is crucial to apply gait analysis in practical settings where cameras are fixed and perspectives are limited, such as surveillance cameras, in contrast to laboratory settings.

1.2 Research questions

The main objective of this study is to proposed a new method for gait analysis in multi-camera environments to overcome the view-variation issue. Since it is one of the challenge for analyze the human gait that can degrade a reliability of the analysis system. The following research questions are set to accomplish our purpose.

- How to analyze human motion from a multi-view gait image for human behavior analysis based on their walking pattern?
- How to improve human gait analysis method from the multi-view gait image sequences for person identification under surveillance scenarios?
- How to explore the optimal feature to estimate the human gait?

1.3 Objectives

According to the research questions as above, the objectives are set to answer it. The following objectives intend to describe the philosophy of this study.

- To analyze the human behavior from the motion based on walking pattern.
- To improve the gait analysis method for person identification from multiple perspectives of surveillance cameras.
- To find the optimal feature for human gait estimation.

Chapter 2

Vision-based gait analysis

This chapter is an explanation of the vision-based gait analysis, starting with an overview. Then, introduces methods, systems, and devices used for an analysis. Followed by the introduction of the gait parameters, the information obtained from the gait, literature review and the motivations and challenges of the gait analysis.

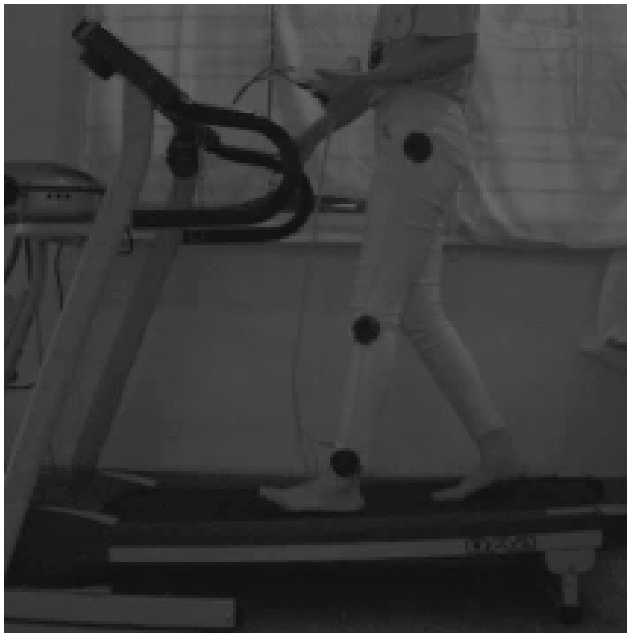
2.1 The overview of vision-based

Vision-based gait analysis proceeds based on the data captured from cameras. It can be a marker-based or non-marker-based analysis. The marker-based is simply a case where we attach markers to the joints of the subject's body and capture their walking sequences. Then, extract the required features based on the marker's position.Figure 2.1a presents a sample image that captures a subject walking on a treadmill with attached markers on her limb.

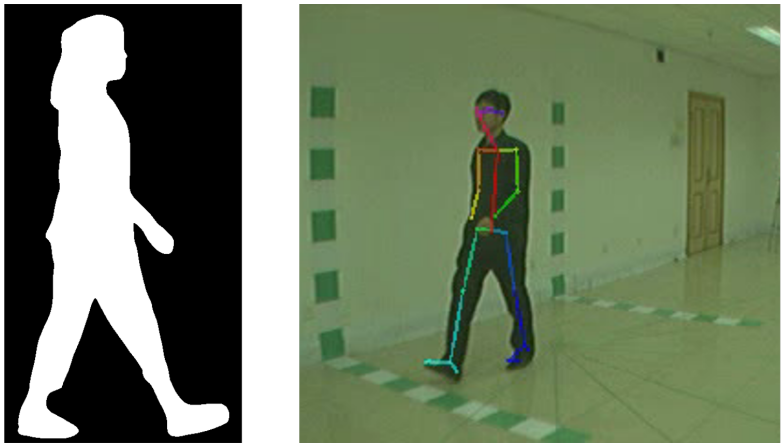
The non-marker-based system requires only cameras to capture the subject's walking sequences. Subsequently, the features will be extracted depending on the requirements of the analysis approach, i.e., appearance-based and model-based, and the required features that will be employed with that approach. Our interest is a non-marker-based system due to it is impossible to attach markers on people in the real-world situation, especially for the surveillance scenario.

Researchers have used a variety of methods to study gait. Here are some examples of Machine Learning (ML) methods:

- Pattern matching
- K-Nearest Neighbor (K-NN)
- Support Vector Machine (SVM)



(a)



(b)

Figure 2.1: Example images to show the different between marker-based and non-marker-based gait analysis. (a) Sample image of the marker-based gait analysis. The markers is attached at the hip, knee, and ankle position to mark interested points for further analysis. (b) Sample image of the non-marker-based gait analysis that processes directly on image with no makers attached on the subject's body. The right image that used for skeleton plot is originally from [55].

As well as Deep Neural Networks (DNNs) methods:

- Convolutional Neural Networks (CNN)
- Long Short-Term Memory (LSTM)

Despite the reliability of both algorithms, this study selects the pattern matching due to the insufficient data for DNNs to learn. Since it is a classification technique, it does not require training state or a large number of datasets.

There are various studies that employ DTW in gait analysis, such as the research of R. Hughes et al. [25]. They improved the floor-based monitoring system and implemented DTW with KNN to enhance walking identification. M. Błażkiewicz et al. [7] applied DTW to assess the gait asymmetry of barefoot walking to evaluate the gait symmetry. The work from Y. Ge et al. [17] employed DTW to match the signals from LoRa sensors with a database to recognize the gait. D. Avola et al. proposed wearable sensor-based gait recognition using a smartphone accelerometer, based on a modified DTW, and applied modified majority voting to return the matched identity of the best comparison score in order to improve the recognition's accuracy [4].

The previous studies show that the pattern matching method is effective in recognizing the gait, and data visualization is possible. However, the DNNs method is crucial for extending the gait analysis beyond recognition tasks, which is our future plan.

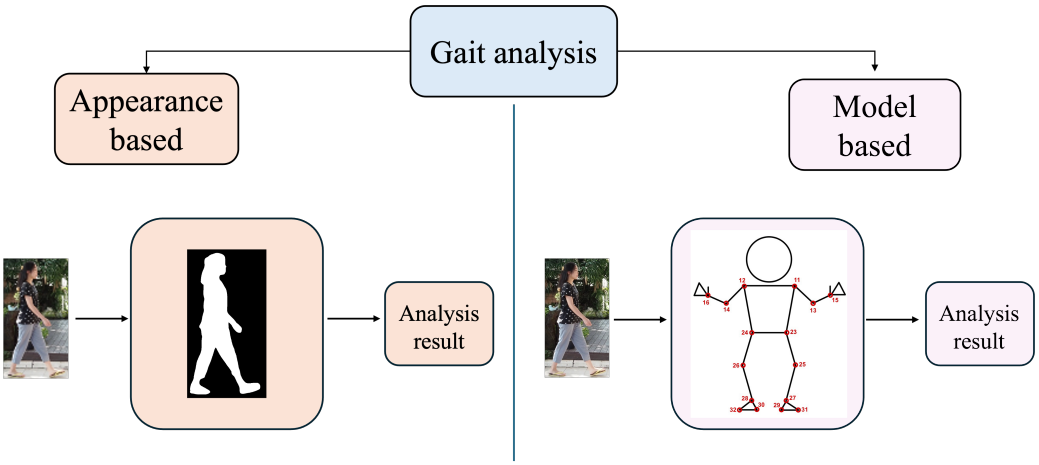


Figure 2.2: Sample visualization of appearance-based and model-based analysis of the gait recognition.

2.2 Approaches to recognize the gait

There are two main approaches for multi-view gait recognition, i.e., appearance-based and model-based approaches [42], as in Figure 2.2.

2.2.1 Appearance-based approach

The appearance-based approach is model-free analysis, which analyzes directly from images or videos and uses shape and textural information as a feature without directly measuring body movements. Various previous studies employed this approach to recognize the gait for identifying people from multiple perspectives from more than one cameras.

The paper from X. Huang et. al [24] employed CNN and used multi-scale features to represent the motions on images. They used Frame-level feature to represents the appearance characteristics, Short-term temporal feature to represents a short period temporal motion patterns, and Long-term feature to represents a combination of the motion from every frames. The work from M. Alotaibi and A. Mahmood [1] intends to increase gait recognition accuracy by developing eight layers of deep CNN that are less sensitive to variations and occlusions. They employed CASIA-B, a multi-view gait database with various walking conditions, for the experiment. Their proposed method can overcome several issues, but the performance will decrease if the gallery set does not cover a variety of walking conditions. C. Fan et al. [13] claim that different parts of the human body consist of diverse visual appearances and movement patterns during walking. GaitPart was proposed as a way to extract gait features. The goal is to improve the learning of part-level features using a frame-level part feature extractor made up of FConv and get the short-range spatiotemporal expression by using a Temporal Feature Aggregator with a Micromotion Capture Module (MCM). GaitEdge is a framework described by J. Liang et al. [32] for recognizing human gait. It can make this framework more practical and keep performance from dropping in cross-domain situations by blocking irrelevant gait information. They designed the module to integrate the trainable edges of the segmented person's shape with the fixed internals of silhouette images based on the mask operation, named Gait Synthesis.

Even if the appearance-based approach achieves great performance in gait recognition with lower system complexity, it requires more complex features for discriminating the gait. The significant challenges of this approach still exist, as it mostly extracts features based on silhouette images, making it sensitive to environmental factors such as lighting and dynamic background, especially when applied to a real-life situation. Additionally, the silhouette

images contain other non-relevant gait information that is unreliable in practice. Furthermore, if parts of the body are occluded, which is mostly self-occlusion, such as crossed legs, it is unable to extract the features. Moreover, it has the inability to capture some features that are invisible through textural and shape, such as joint angles. These reasons make the appearance-based approach suitable for the controllable setting environment, such as a laboratory setting, unlike the real situation, which includes various uncontrollable conditions.

2.2.2 Model-based approach

The model-based approach requires a mathematical model to distinguish the gait characteristics. This approach requires a prior human model, such as the human pose estimation algorithm that is used to extract the joint coordinates, unlike the appearance-based approach, which employs features directly from images. It makes a model-based approach more robust to the surrounding environmental conditions and allows us to extract various features in addition to shape and textural information.

As in the previous studies from R. Liao et al. [34]. They proposed a model-based gait recognition by extracting 14 body joints of 2D human pose estimation from images and transforming them into 3D poses, called PoseGait. The CNN is implemented to learn the handcrafted gait features which is a fusion of 3D poses, joint angle, limb length, and joint motion. They evaluate the proposed method by a cross-view recognition on the CASIA-B, a multi-view gait database. T. Teepe et al. conducted a gait recognition based on the skeleton landmark of human joints from a Graph Convolutional Networks (GCNs) named GaitGraph2 [50]. They used pre-calculated joint positions, motion velocities, and bone features that extracted from the skeleton-based information, and implemented the ResGCN architecture to construct the model for gait recognition. Their research focus on the useful and reliable gait features for further study that aim to apply with a practical situation. They apply the multi-view gait database, the CASIA-B and OUMVLP-Pose to evaluate the work and achieve outstanding performance on OUMVLP-Pose dataset. K. Han et al. proposed a discontinuous gait image recognition based on the extracted keypoints of the human skeleton [21]. They aim to overcome the situation of discontinuity in the gait images. This study achieves a high recognition rate and is robust to common variations. Y. Fu et al. proposed a frame work to generalized the model-based approach for gait recognition [16]. They focused conducted a preliminary study and found that previous studies were lack of the important issue that caused the degraded of recognition performance when performed with un-

seen scenarios, which is a generalization of joint keypoints. They proposed Human-Oriented Transformation (HOT) and Human-Oriented Descriptors (HOD). The HOT used for transforming the skeleton sequences in camera coordinate system into human-oriented coordinate system, and implemented HOD to obtained the features based on body ratio and structure. Moreover, they designed a Part-Aware Graph Convolutional Network (PAGCN) to learn the relationship between features. They evaluated the work by recognizing on the same dataset (source-domain) and across the different dataset (cross-domain), and achieves great performance on cross-domain without decreased the performance on source-domain testing.

Compared to the appearance-based approach, the model-based approach loses shape information and requires a more complex system to address the human body's skeleton landmark for obtaining joint coordinates and employing it for further feature extraction. However, the features used in the model-based approach are more simple yet have strong potential to achieve impressive results compared with the appearance-based approach, in which the system is less complex but requires more complicated features to discriminate the gait. Moreover, the gait features from the model-based approach are possible to apply in the real situation because they relate directly to the motion of the human body and are robust to the surroundings. Meanwhile, the features from silhouette images include the irrelevant gait information that can mislead the system to the wrong analysis. On top of that, appearance-based suffers with the occlusions, especially a self-occlusion, because it rely on silhouette images that is enriched with shape and textural information.

Previous studies show that the model-based approach for vision-based gait analysis primarily extracts features based on joint coordinates from the pose estimation algorithm. It suggests that our proposed method is a model-based approach.

Mostly, the existing studies applied DNNs to the gait analysis and tried to recognize the known persons with the unknown persons. However, our proposed method aims to identify the people we already know and ignore the others, for example, searching for the suspicious person. Thus, the DNNs are unnecessary for this purpose, the pattern matching can perform this task. Additionally, the previous studies tried to recognize people across different perspectives to overcome the view-variation by DNNs, but we employ the voting algorithm to integrate the information from multiple cameras to overcome the view-variation.

2.2.3 Vision-based human pose estimation

In recent years, vision-based joint estimation has been widely deployed to extract the human joint landmark for vision-based gait analysis, especially for the model-based approach, which is simpler to implement. It requires no additional cost or time spent on equipping markers, only setting the scene with cameras is required. There are various pre-trained human pose estimation models that can be applied. This makes the model-based approach more accessible and affordable.

This section presents an example of features based on joint parameters. Let $\mathbf{p}^{j,i,t} = [x, y]^T$ represent a location of joint j on x-axis and y-axis in Figures 2.3a and 2.3b, which present the human body joints from MediaPipe [5] and OpenPose [8], respectively. The parameters i represent a person index, and t represents time or frame number. The changing of $\mathbf{p}^{j,i,t}$ can present a walking pattern. By this definition, a walking pattern includes not only a lower body but also the upper body. When a person walks, their entire body moves, leading to the correlation of all joints.

There are various state-of-the-art techniques for vision-based human pose estimation, such as MediPipe [5], OpenPose [8], and AlphaPose [15], [29], [14]. Researchers have applied these most commonly used pose estimations to various human-related fields such as activity, gait, hand gestures, and facial recognition, as some benchmarks offer detailed estimations of face and hand landmarks.

- *openpose pose estimation*

  Openpose is a real-time multi-person pose estimation based on Part Affinity Fields (PAFs) that include face, hand, body, and foot landmarks [8], [41], [9], [51]. It is a bottom-up approach that begins with locating the position and orientation of the limb in an image, followed by estimating the other parts.

  OpenPose, the most widely used pose estimation tool, supports research purposes by providing an open-source library compatible with many platforms. Moreover, it can apply to both the CPU and GPU to run the program, depending on the model. There are three human landmark models that can be employed with Openpose, i.e., MPII [3] that produces 15 keypoints of human joints, COCO [35] that returns 17 keypoints of human joints, and body 25 models, which is a COCO with extended feet landmarks that gives 26 keypoints of human joints.

  OpenPose is applicable to extracting the 3D keypoints with the requirement of the stereo cameras. It is unable to extract 3D keypoints with a single camera.

- *AlphaPose pose estimation*

  AlphaPose is an open-source system for a real-time multi-person pose estimator [15], [29], [14]. It is similar to OpenPose, but the difference between them is the estimation method. AlphaPose is a top-down strategy that detects the human first and then estimates their pose. Furthermore, it allows pose tracking, which predicts the pose keypoint over time in the sequences to separate multiple people's identities. The COCO model [35] with 17 joints keypoint and the Fast Pose model with 26 joints keypoint.

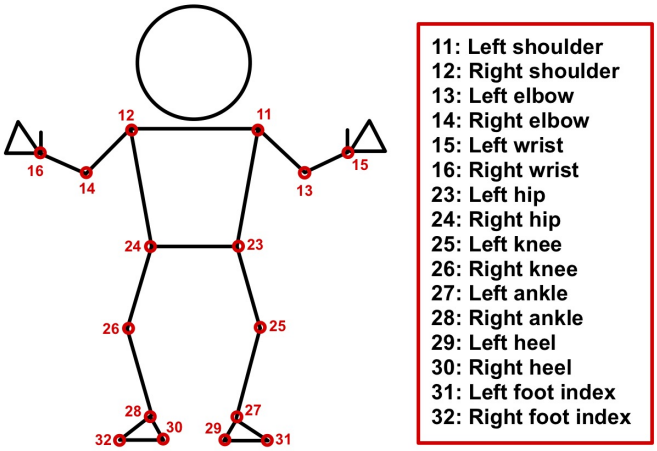
  Accordingly, it is an open-source system and supports both the Windows and Linux platforms. In addition, it provides an online platform for open-source pose trackers.

- *MediaPipe pose estimation*

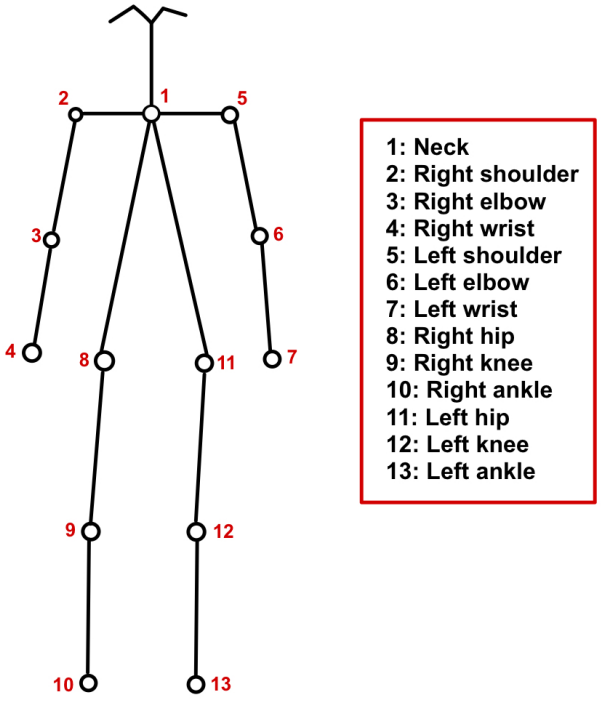
  MediaPipe is an open-source single-person pose estimation tool from Google. It uses the BlazePose model [5], which is a lightweight ML model that produces 33 keypoints of human landmarks, including hands and faces, but we employed only 16 keypoints for the experiment, as shown in Figure 2.3a. Each keypoint contains the coordinates in the x, y, and z axes. Moreover, it provided joint coordinates for both image coordinates and real-world coordinates. The real-world coordinates present the x, y, and z in a unit of meters, where z is a depth.

There are previous studies that show comparative results between three benchmarks. X. Li et al. [31] proposed fitness action counting and classification based on MediaPipe. They present the comparative results between MediaPipe, OpenPose, and AlphaPose, which claim that MediaPipe is faster to recognize and achieves high accuracy. K. Y. Chen et al. also used MediaPipe to get the features they needed to use transfer learning deep neural networks to find the type of fitness movement and how complete it was [10]. They also suggested that MediaPipe has an uncomplicated implementation, fast computational speed, and high accuracy.

Additionally, we check the the self-occlusion on MediaPipe by selecting a sequence that includes the overlap of two legs, as shown in Figure 2.4a. Then, we plot the skeleton of joint coordinates on selected image as shown in Figure 2.4b, as well as the extracted coordinates and confident score. It suggests that the pose estimation algorithm can handle the self-occlusion problem even though some parts are occluded, e.g., left arm and left knee.

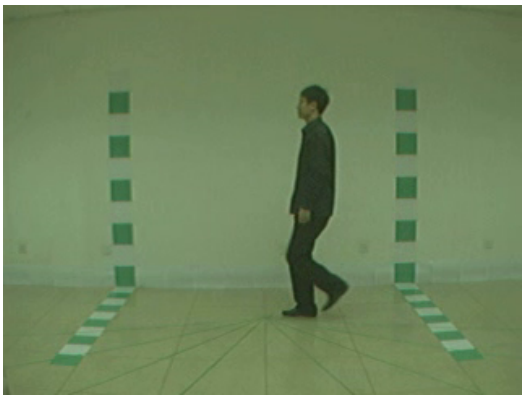


(a)

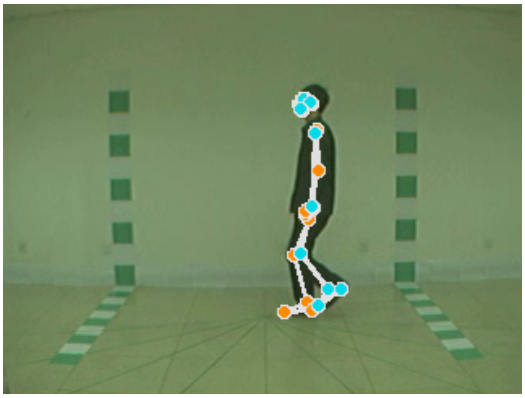


(b)

Figure 2.3: Human pose landmarks that used in this study. (a) 16 keypoints from MediaPipe [5] pose estimation. (b) 13 keypoints from OpenPose pose [8] estimation.



(a)



(b)

Figure 2.4: Self-occlusions problem testing by MediaPipe on a sequence from CASIA-B (90°) [55]. (a) Original image and (b) Skeleton plot.

2.3 Gait information

Gait is a periodic sequence of joint movement in which each joint movement pattern is repeated as cycles. In a cycle, it consists of a change in joint position over time. When joints have motion, postures occur. Each person has different postures according to their individual walking speed, arm swing, foot placement, weight transfer, etc. These behaviors refer to the neurological control that expresses our individual walking trait, which represents the walking pattern. It is noticeable when we visualize a human landmark frame-by-frame as a sample in Figure 2.5.

Equation (2.1) defines the changing of joints location when walking over time, which can represent the walking pattern. We call it 'posture' because it records human postures while walking frame-by-frame.

$$\text{posture} = \begin{bmatrix} [\mathbf{p}^{j_1,i,0} ... \mathbf{p}^{j_u,i,0}] \\ \vdots \\ [\mathbf{p}^{j_1,i,t} ... \mathbf{p}^{j_u,i,t}] \\ \vdots \\ [\mathbf{p}^{j_1,i,n} ... \mathbf{p}^{j_u,i,n}] \end{bmatrix} \tag{2.1}$$

The $\mathbf{p}^{j_u,i,t}$ represents the location of joint j_u, where $u \in \mathbb{U}$ and $\mathbb{U}$ is a number of joints, i is a person index, and $t = 0$ and $t = n$ represent the first and last frames in a sequence, respectively.

Furthermore, it can present information about the transportation mode, such as the direction, and use it to predict the walker's destination. It also includes individual information that represents the walker's personality, medical condition, behaviors, and emotions. We can extract the mentioned information by distance in any direction on horizontal, e.g., front, rear, side, and also on multi-vertical view, such as on the top. Since gait is a walking trait, it is hard to pretend, copy, or change.

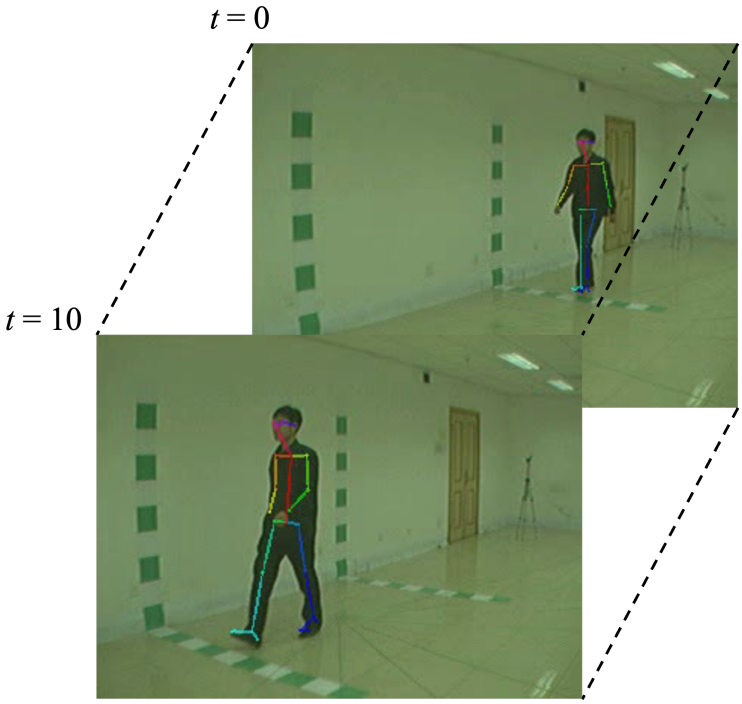


Figure 2.5: Sample images to demonstrate how joints are changed its position over times from t=0 to t=10. Where t represents time or frame number in this study. Images used for demonstration are from CASIA-B dataset [55]

2.4 Motivations & Challenges

According to the previous studies have addressed that the problem with multi-view gait analysis was that the gait information would change when the camera angle changed, leading to less reliable analysis results. Our aim is to overcome this problem and improve multi-view gait analysis in surveillance scenes. As the reliable identification of video surveillance cameras can improve security. By identifying suspicious individuals through their gait, we can prevent them from changing their appearance to avoid detection.

Chapter 3

Gait analysis in surveillance scenes

This chapter describes the gait analysis in surveillance scenes based on joint features. We start by working on single-view gait images before extending to a multi-view gait analysis. It includes the definition of a single-view scenario and discussions on a single-view gait analysis. The next step is to introduce multi-view gait analysis, which builds upon single-view gait analysis and outlines the methodology for analyzing the multi-view gait images. This section includes an explanation of the methods, equations, and calculations to clarify our methodology that was deployed in this study, which is a major part of this research.

3.1 Single-view gait analysis

A single-view scenario involves capturing a scene from a single perspective using a single camera. Figure 3.1 shows a sample setting environment to obtain single-view scenarios, where a single camera captures the walking within its field of view. This setup has served as the foundation for vision-based gait analysis in the past years. Its simplicity and accessibility have made it a common starting point for various research projects in this field. Furthermore, it avoids the difficulties related to view variation because it is a single perspective. However, the perspective and coverage area of a single camera limit the information it can capture. These make the extracted information useful in that identical perspective, which is not covered when changing the perspective or camera angle.

Our aim is to properly understand the concepts of gait analysis. Therefore, we begin this research by investigating single-view gait analysis, as prior

studies have done, before continuing to multi-view gait analysis.

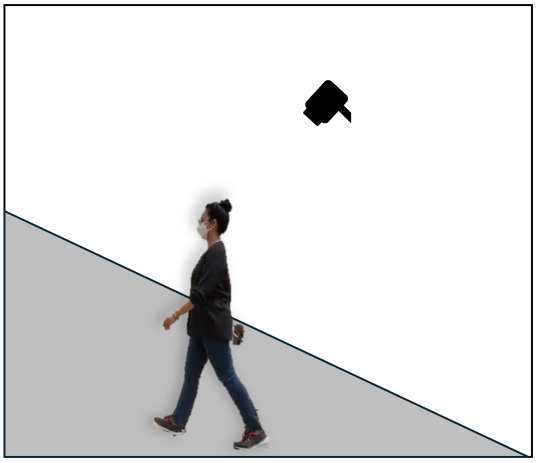


Figure 3.1: Sample of a single-view setting environment to obtain single perspective sequences in surveillance scenario.

3.1.1 Patterns matching based on Cosine similarity

We begin with a single-view gait analysis using Euclidean distance to measure the relative distance between two joints on the lower body of a human landmark, estimated by MediaPipe. Usually, the definition of the gait refers to the lower body, including the hip, knee, ankle, and feet, as defined by J. Perry and J. M. Burnfield [38]. Therefore, our focus is on the changes between the lower body parts. Then, we apply cosine similarity to match the patterns.

Figure 3.2 presents sample images of a dataset used in a single-view gait analysis experiment, which was collected by the authors. It is a walking scene with eight subjects with different backgrounds.

The Euclidean distance is defined as the length of the line segment between two points. We use it to measure the distance from the first point to the second point. We obtain the joint coordinates by using MediaPipe to estimate the human pose landmarks. We calculate the Euclidean distance between these joints as shown in Figure 3.3, i.e., the distance between hip, knee, ankle, heel, and toe, to extract their periodic patterns. Equation 3.1 presents a calculation of Euclidean distance between two interested joints, $\mathbf{p}^{j,i,t,D}$ and $\mathbf{p}^{(j+2),i,t,D}$.

$$ed = \|\mathbf{p}^{j,i,t,D} - \mathbf{p}^{(j+2),i,t,D}\| \tag{3.1}$$

Finally, we apply cosine similarity to measure the parity between patterns. Since it is a similarity measurement between two vectors, $\mathbf{ED}_1$ and $\mathbf{ED}_2$ represent the vectors that store Euclidean distance along the time t of the sequences 1 and 2, respectively, as described in Equation (3.2), and Equation (3.3) presents the way to score the similarity between two patterns.

$$\mathbf{ED} = \left[ed^0 ... ed^t ... ed^n\right] \tag{3.2}$$

$$\text{Cosine similarity}(\mathbf{ED}_1, \mathbf{ED}_2) = \frac{\mathbf{ED}_1 \cdot \mathbf{ED}_2}{|\mathbf{ED}_1| \cdot |\mathbf{ED}_2|} \tag{3.3}$$

Our findings from this experiment indicate that these separated features are insignificant for identity matching. Even though we extracted a pattern of the distance between two joints that changes over time, when humans walk, all body parts have movement. Thus, we should employ the features together, not individually. Additionally, it suggests we employ body angles instead of the relative distance between two joints. According to an observation from frame-by-frame skeleton landmark images, as shown in Figure 2.5, it presents the way joint angles occur when humans walk and the walking posture is changed over time. Hence, we focus on employing whole-body joint angles instead.

Moreover, the cosine similarity is improper to match the time series data that vary timing and speed, that is, when the corresponding points in two sequences may not line up perfectly. The cosine similarity has its own advantages and is suitable for comparing the similarity of vectors in high-dimensional spaces. However, it may not be appropriate for time series data where temporal relationships and variations need to be considered.

3.1.2 Patterns matching based on Dynamic Time Warping (DTW)

After we acknowledged that the Cosine similarity is inappropriate to discriminate the walking patterns, the Dynamic Time Warping (DTW) is then applied according to the time warping characteristic that can handle the time series data.

We extracted the joint angles based on the joint coordinates that were extracted from MediaPipe. According to the paper from M. Alvaro et al. [36], the joint angle is one of the most commonly used features for gait analysis.

Figure 3.2: Sample images of the dataset used in a single-view gait analysis.

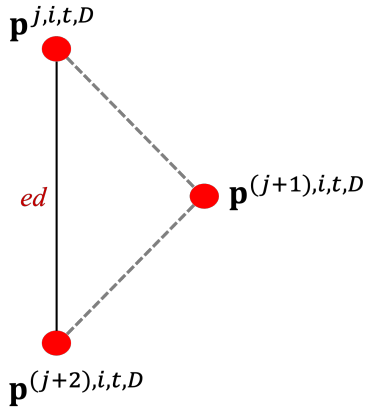


Figure 3.3: Euclidean distance between two joints description.

Additionally, the other features for gait analysis can be extracted from joint angle. Hence, the joint angle is selected as a feature in this study.

Furthermore, W. Pirker and R. Katzenschlager [39] summarized the important parameters for clinical examination in their paper in addition to the listed parameters above, such as arm swing. It is crucial to know that gait analysis involves the whole body. Therefore, to enhance our understanding of an individual's gait, we need to concentrate on the entire body.

To extract the joint angles, we initialize the process by connecting three joints, i.e., $\mathbf{p}^{(j-1),i,t,D}$, $\mathbf{p}^{j,i,t,D}$, and $\mathbf{p}^{(j+1),i,t,D}$, from the human pose estima-

tion landmarks in Figure 2.3a as a triangle. Figure 3.4 represents the mentioned triangle connection. We determine the Euclidean distance between each joint to generate a triangle. We let it be Legs a, b, and c using Equations (3.4)–(3.6). Leg a represents a connection line between joints $\mathbf{p}^{(j-1),i,t,D}$ and $\mathbf{p}^{(j),i,t,D}$, leg b is a connection line between joints $\mathbf{p}^{j,i,t,D}$ and $\mathbf{p}^{(j+1),i,t,D}$, and leg c is a connection line between joints $\mathbf{p}^{(j-1),i,t,D}$ and $\mathbf{p}^{(j+1),i,t,D}$. Finally, we apply the cosine law in Equation (3.7) to extract the middle angle ($\theta^{j,i,t,D}$), which is a preferred joint angle to use as a feature.

$$a = \|\mathbf{p}^{j,i,t,D} - \mathbf{p}^{(j-1),i,t,D}\| \tag{3.4}$$

$$b = \|\mathbf{p}^{j,i,t,D} - \mathbf{p}^{(j+1),i,t,D}\| \tag{3.5}$$

$$c = \|\mathbf{p}^{(j-1),i,t,D} - \mathbf{p}^{(j+1),i,t,D}\| \tag{3.6}$$

$$\theta^{j,i,t,D} = \cos^{-1}\left(\frac{(a^2 + b^2) - c^2}{2 \times (\sqrt{a^2} \times \sqrt{b^2})}\right) \tag{3.7}$$

In this research, we extracted ten angles, including elbow, hip, knee, front ankle, and back ankle (both left and right sides) from using MediaPipe as a pose estimation method. Thus, after obtained the middle angle as $\theta^{j,i,t,D}$, the $\boldsymbol{\theta}^{i,t,D}$ variable in Equation (3.8) represents a feature vector that used for pattern matching, and we gathering it together as vector of (10,1) dimension.

$$\boldsymbol{\theta}^{i,t,D} = \left[\theta^{1,i,t,D} ... \theta^{j,i,t,D} ... \theta^{10,i,t,D}\right]^T \tag{3.8}$$

Table 3.1: Results of DTW distance on the dataset in Figure 3.2

		True label							
		0	1	2	3	4	5	6	7
Matched	0	253.12	305.36	324.22	507.30	254.47	273.08	481.51	285.35
	1	320.37	**251.15**	330.81	**447.17**	282.05	311.17	516.14	255.89
	2	272.33	305.84	265.60	485.96	267.31	296.93	469.75	265.90
	3	297.48	298.39	303.35	505.97	254.76	268.39	507.73	288.14
	4	296.69	306.25	283.11	516.98	239.55	307.05	470.09	290.69
	5	263.28	290.57	296.52	518.08	284.77	**196.10**	452.85	263.81
	6	**214.67**	285.65	**263.33**	500.85	246.70	271.27	**448.29**	**224.86**
	7	308.87	269.86	324.70	504.57	**237.04**	308.64	504.14	248.21

Table 3.1 presents the calculated DTW distance of the dataset shown in Figure 3.2. The single view accuracy of the correct matching from using DTW distance is 42.90%. From this experiment, we found that DTW is suitable to handle the time series data and produces great result on this dataset. Unfortunately, a single-view is unable to recognize the gait in the other perspectives. According to this limitation, we move our focus to multi-view gait analysis.

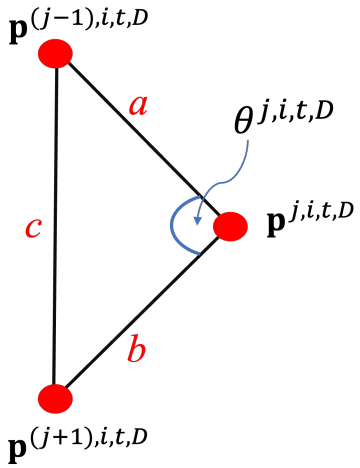


Figure 3.4: Joint angle calculation using cosine law to calculate a middle angle ($\theta^{j,i,t,d}$) between 3 joints.

3.2 Multi-view gait analysis

Multi-view scenario is a scene captured by more than one camera with different perspectives. Figure 3.5 shows a sample of the setting environment using three cameras to obtain scenarios with different perspectives. The laboratory setting usually equips multiple cameras consecutively to capture the same scene as a sample in Figure 3.6a. Figure 3.6b presents the obtained images from the setting environment in Figure 3.6a. In practice, this scenario equips multiple cameras separately and may be non-consecutive, resulting in more challenges to an analysis of the multi-view gait images.

The view-variation problem is one of the significant challenges of multi-view gait analysis. This issue arises because, as the perspective changes across different camera angles, the classified identity may vary significantly. Even if it is the same person walking, their gait can be recognized as that of a different person according to this variation. Many researchers proposed algorithms that aimed to handle the view-variation problem by employing computer vision, machine learning, and deep learning methods. However, the view-variation problem remains significant in multi-view gait analysis as it decreases the understanding of human movement across different perspectives, especially in real-world scenarios such as scenes from a surveillance camera.

In this study, we apply a pattern matching to match walking patterns for person identification purposes. Moreover, we apply a majority vote to address the view-variation issue. This strategy aims to aggregate the information from multiple perspectives to enhance the accuracy and reliability of the pattern matching technique over the view-variation problem.

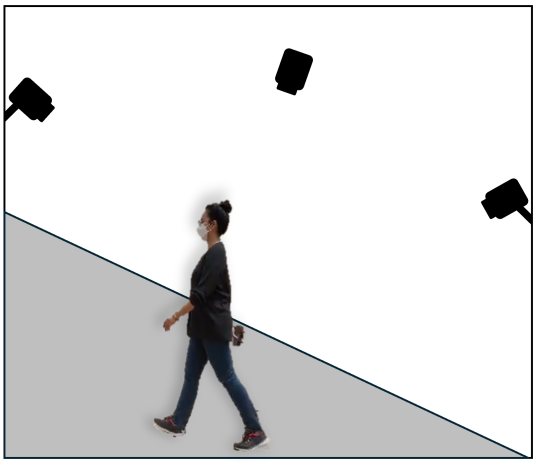
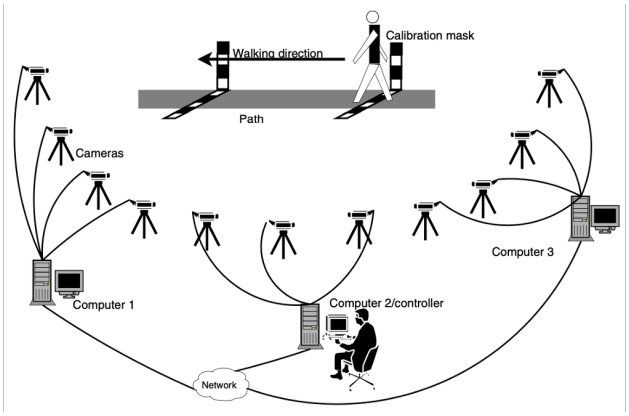


Figure 3.5: Sample of a multi-view setting environment from.



(a)

(b)

Figure 3.6: Sample of a multi-view in laboratory setting environment images from [55]. (a) Setting environment of the multi-view sequences. (b) Obtained scenarios of multi-view sequences.

3.3 Methodology for multi-view gait analysis

Figure 3.7 shows our overall methodology for this study. We separate the sequences into reference, which is our database, and target, which is the input sequence to be analyzed. Each sequence is a captured person i walking scenario from each camera perspective (D).

The authors extract the joint coordinates of the reference and target sequences as $\mathbf{p}^{j,i_k,t,D}$ and $\mathbf{p}^{j,i_{tar},t,D}$. The variable k denotes the number of reference sequences, which is $k \in \mathbb{K}$. We let j be a joint number, D be a camera perspective, and i be a person identity label. Then, we extract the features from reference and target sequences as the joint angles ($\boldsymbol{\theta}^{i_{ref},t,D}$ and $\boldsymbol{\theta}^{i_{tar},t,D}$). The joint angles extracted from MediaPipe keypoints consist of ten angles, including the elbow, hip, knee, and ankle (front and back). The joint angles extracted from OpenPose keypoints consist of six angles according to the feet landmarks are not included. Thus, the extracted joint angles consist of the elbow, hip, and knee. We extract each joint angle from both the left and right sides of the body.

Next, we calculate the correlation between joint angles of reference and target sequences as $\mathbf{c}^{i_{ref},t,D}$ and $\mathbf{c}^{i_{tar},t,D}$, which represent the frame-by-frame or time-dependent correlation feature. Additionally, we calculate the rank correlation between each joint angles from the entire sequences to represents the overall individual walking pattern as $\mathbf{c}^{i_{ref},D}$ and $\mathbf{c}^{i_{tar},D}$. In the other word, it is a time-independent correlation feature.

Then, we apply the distance calculation to determine the distance between reference and target sequences, which will divided into two approaches based on the extracted features.

- Approach 1 & 2: Apply Dynamic Time Warping (DTW) with time-dependent features.

For this approach, we apply DTW to match the patterns based on the time-dependent features, which are joint angles (Approach 1) and time-dependent correlation features (Approach 2), as shown in the diagram in Figure 3.8a. The reason is that these features, including time information, require the time warping characteristic of DTW to match the patterns that vary in speed, time, and length. We let the measured DTW distance be $S^{i,D}_{\text{DTW}}$.

- Approach 3: Apply Euclidean distance (EU) with time-independent feature.

Essentially, the correlation requires variations of data in order to see their relationship and association. We then propose the feature representative

to represent the walking pattern of the entire sequence. For this reason, time information will be ignored. Hence, direct matching is applied in this approach.

We proposed another approach to studying the gait by calculating the time-independent correlation, as shown in Figure 3.8b. The rank correlation between each joint angles will be calculated, which will re-arrange the entire sequential data by its ranking order. This makes the time-independent correlation the feature descriptor that ignores the time order, which implies that the variation of time, speed, and data length disappear. Therefore, we can apply the Euclidean distance (EU) to this feature. We let the calculated EU be $S_{\mathrm{EU}}^{i,D}$.

After determining the distance, we find a minimum distance to match a person's identity in the target sequence with the reference sequence. Then, the matched person identity is returned as i_k^D, which refers to the person identity in each camera perspective (D). Finally, we aggregate the separated person identity (i_k^D) from each D by applying majority voting to increase the reliability of a person identity matching.

In fact, we select the unsupervised learning method, i.e., DTW and EU matching. However, the definition of unsupervised learning is clustering the input data without having labeled data to supervise them. Meanwhile, the supervised learning method will classify the input data according to the labeled data, such as the database, in which the K-NN is included. Our method is pattern matching that includes labeled identity in the reference (or database), which should be categorized as the supervised method.

Unfortunately, both our method and K-NN are not learning anything from the features or input data. Even though its category is the supervised learning method, it is not included in DNNs, but it is a machine learning. The different between ours and K-NN is we perform a pattern matching across every samples in the identical view database, but K-NN perform a matching across every sample in database, includes the different view.

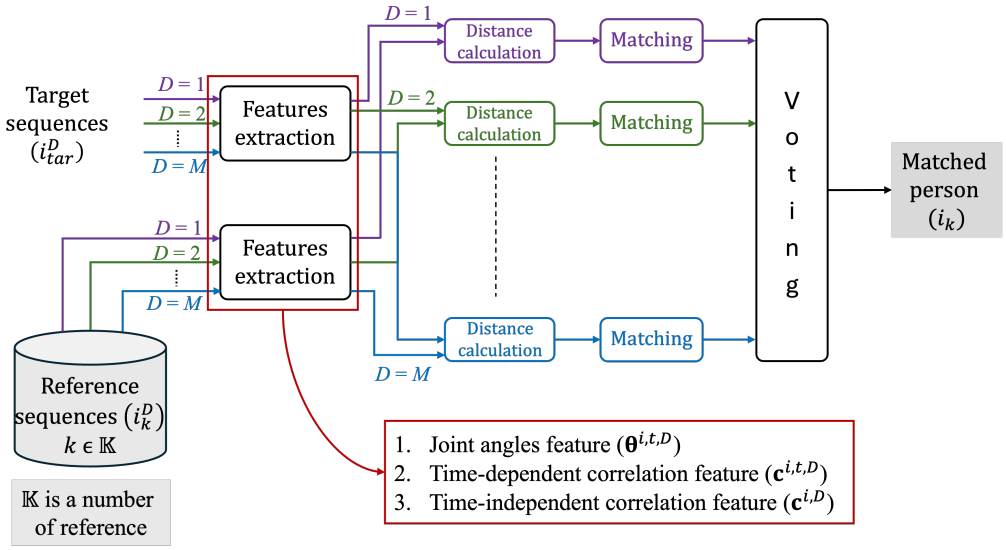
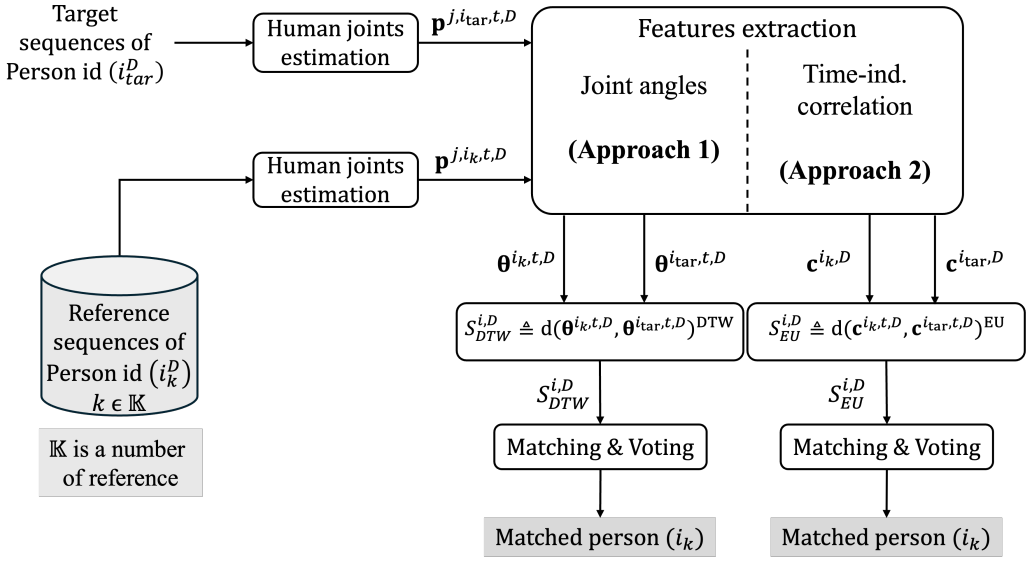
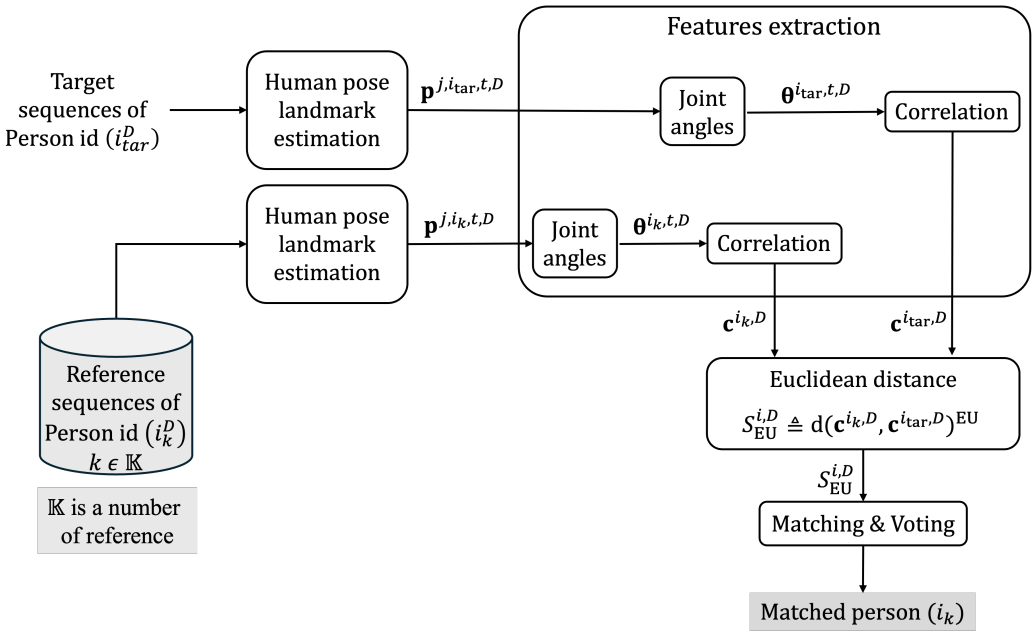


Figure 3.7: Overall methodology for walking pattern matching in this study.



(a)



(b)

Figure 3.8: Diagram of two approaches for walking pattern matching based on features. (a) Approach 1 & 2: Apply Dynamic Time Warping (DTW) with time-dependent features. (b) Approach 3: Apply Euclidean distance (EU) with time-independent feature.

3.4 Calculations & equations

3.4.1 Features extraction

- *Joint angles calculation*

The procedure of joint angles extraction is the same as described in single-view gait analysis. We apply law of cosine in Equation (3.7) to extract the middle angle ($\theta^{j,i,t,D}$) of a triangle that perform by connecting three joints, i.e., $\mathbf{p}^{(j-1),i,t,D}$, $\mathbf{p}^{j,i,t,D}$, and $\mathbf{p}^{(j+1),i,t,D}$ together. We calculate the three legs of a triangle by calculating the Euclidean distance between two points as Equations 3.4–3.6, and let it be leg a, leg b, and leg c. Finally, the desired angles are gathered in a vector form as an Equation 3.9.

This experiment separates feature vectors into three parts, i.e., whole, upper, and lower body, to determine the distance for a matching purpose. Typically, human gait refers to the motion of lower body parts, i.e., the hip, knee, and ankle. However, we notice that the whole body has motion while humans walk, not just the lower parts. Thus, we decide to employ the upper body feature to study the effect of the body parts on an analysis of the gait.

The $\boldsymbol{\theta}^{i,t,D}$ variable in Equation (3.9) represents a feature vector of the whole body. Upper body ($\boldsymbol{\theta}_u^{i,t,D}$) consists of two angles, left and right elbow, as in Equation (3.10), and the lower body ($\boldsymbol{\theta}_l^{i,t,D}$) consists of the remaining angles as in Equation (3.11). Where $u \in \mathbb{U}$ represents the number of joints. The $\mathbb{U} = 10$ joints for MediaPipe, and 6 joints per each for OpenPose and AlphaPose.

$$\boldsymbol{\theta}^{i,t,D} = \left[\theta^{1,i,t,D} ... \theta^{j,i,t,D} ... \theta^{u,i,t,D}\right]^T \tag{3.9}$$

$$\boldsymbol{\theta}_u^{i,t,D} = \left[\theta^{1,i,t,D} \theta^{2,i,t,D}\right]^T \tag{3.10}$$

$$\boldsymbol{\theta}_l^{i,t,D} = \left[\theta^{3,i,t,D} \theta^{4,i,t,D} ... \theta^{j,i,t,D} ... \theta^{u,i,t,D}\right]^T \tag{3.11}$$

- *Time-dependent correlation calculation*

The time-dependent correlation calculation used to calculate the pearson correlation between each pair of joint angles of the same t, or it is a frame-by-frame correlation that proposed as frame-by-frame feature descriptor of the walking pattern. Since the correlation calculation requires at least two values per a variable, so it is impossible to calculate the correlation between each $\theta^{j,i,t,D}$. Hence, the time-dependent correlation will be the correlation between each pair of joint angles.

Table 3.2: Sample of the G and H that store the values of elbow angles and hip angles to be used for calculating the correlation between them.

G	H
$\theta^{1,i,t,D}$	$\theta^{2,i,t,D}$
$\theta^{3,i,t,D}$	$\theta^{4,i,t,D}$

The pearson correlation uses to measure the linear relationship between two variables, we let it be G and H. The calculated correlation is $[-1, 1]$, which implies that two variables are similar and have a negative or positive linear relationship. Table 3.2 shows an example of the G that stores the values of elbow angles ($\theta^{1,i,t,D}$ and $\theta^{2,i,t,D}$), and H that stores the values of hip angles ($\theta^{3,i,t,D}$ and $\theta^{4,i,t,D}$). Each G and H store the value of a pair of joint angles that represent left and right parts of each angle.

Next, we calculate the correlation between each pair of joint angles by using Equation 3.12.

$$c_{GH}^{i,t,D} = \frac{\mathrm{E}[GH] - \mathrm{E}[G]\mathrm{E}[H]}{\sqrt{\mathrm{E}[G^2] - \mathrm{E}[G]^2}\sqrt{\mathrm{E}[H^2] - \mathrm{E}[H]^2}} \tag{3.12}$$

Then, the final form of time-dependent matrix ($\mathbf{c}^{i,t,D}$) is shown in Equation 3.13. It stores all $c_{GH}^{i,t,D}$ of v pair of joint angles, where $v \in \mathbb{V}$ represents the number of pair of joint angles.

$$\mathbf{c}^{i,t,D} = \begin{bmatrix} c_{G_1H_1}^{i,t,D} & \cdots & c_{G_1H_v}^{i,t,D} \\ \vdots & \ddots & \vdots \\ c_{G_vH_1}^{i,t,D} & \cdots & c_{G_vH_v}^{i,t,D} \end{bmatrix} \tag{3.13}$$

- *Time-independent correlation calculation*

In this study, we calculate the time-independent correlation between joint angles by implementing rank correlation. Based on frame-by-frame human pose extraction, we extract individual joint angles with respect to the frame (or time t), resulting in a pattern that is time-dependent. The rank correlation aims to be a time-independent feature that can be used as a feature descriptor for the overall individual's pattern.

Spearman correlation is a method to measure dependence between two ranking variables [46]. It is a non-parametric rank measurement that employs a monotonic function to define a relationship between them. The calculated correlation is $[-1, 1]$, which implies that two variables are similar and have a positive monotonic relationship when it is closer to 1. However, if it is closer

to -1, the two variables are perfectly opposite and have a negative monotonic relationship. If a calculated value is close to 0, there is no correlation.

Figure 3.9 shows the way to assign the ranks to $\theta^{1,i,t,D}$ and $\theta^{2,i,t,D}$ values. The $\theta^{j,i,t,D}$ value represents values of joint angle j in time t. Figure 3.9a presents the values of $\theta^{1,i,t,D}$ and $\theta^{2,i,t,D}$ before assigning the ranks, and Figure 3.9b presents the rankings of $\theta^{1,i,t,D}$ and $\theta^{2,i,t,D}$ as X and Y, respectively. The lowest rank value is assigned to the maximum value, and the highest rank value is assigned to the minimum value. Then, arrange the assigned ranks from highest rank to lowest rank values. For the tied ranks, the average number between them will be assigned to all tied ranks. Figure 3.9a shows that there are two identical values of $\boldsymbol{\theta}^{2,i,D}$, which actually are orders of 3 and 4, but we assign an order of 3.5 as they are tied ranks.

t	$\theta^{1,i,t,D}$	$\theta^{2,i,t,D}$
0	149	144
1	148	140
2	146	140
3	145	141
4	143	139

(a)

X	Y
5	5
4	3.5
3	3.5
2	2
1	1

(b)

Figure 3.9: Sample ranking of $\theta^{1,i,t,D}$ and $\theta^{2,i,t,D}$ for calculating the correlation between them. (a) is the values before ranking of $\theta^{1,i,t,D}$ and $\theta^{2,i,t,D}$. (b) is the values after ranking of $\theta^{1,i,t,D}$ (X) and $\theta^{2,i,t,D}$ (Y).

The Equation (3.14) employs to calculate the correlation between two joint angles as $c_{XY}^{i,D}$, where the E[•] value represents an expected value.

$$c_{XY}^{i,D} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - \mathrm{E}[X]^2}\sqrt{\mathrm{E}[Y^2] - \mathrm{E}[Y]^2}} \tag{3.14}$$

Finally, all $c_{XY}^{i,D}$ between u joint angles are stored in a matrix as shown in Equation 3.15. The final form of a matrix that stores correlation between each joint angles of the entire sequences are shown as an example in Table 3.3.

$$\mathbf{c}^{i,D} = \begin{bmatrix} c_{X_1Y_1}^{i,D} & \cdots & c_{X_1Y_u}^{i,D} \\ \vdots & \ddots & \vdots \\ c_{X_uY_1}^{i,D} & \cdots & c_{X_uY_u}^{i,D} \end{bmatrix} \tag{3.15}$$

Table 3.3: Sample of the calculated individual correlation between each joint angle.

	LElbow	RElbow	LHip	RHip	LKnee	RKnee	LAnkle (Front)	RAnkle (Front)	LAnkle (Back)	RAnkle (Back)
LElbow	1.00	0.43	-0.47	-0.32	-0.65	-0.30	-0.31	-0.01	0.10	0.03
RElbow	0.43	1.00	-0.37	0.06	-0.19	0.17	-0.34	-0.46	0.24	0.54
LHip	-0.47	-0.37	1.00	0.06	0.58	0.28	0.07	-0.08	-0.23	-0.30
RHip	-0.32	0.06	0.06	1.00	0.33	0.86	-0.33	-0.54	0.04	0.45
LKnee	-0.65	-0.19	0.58	0.33	1.00	0.47	0.30	-0.34	-0.52	0.08
RKnee	-0.30	0.17	0.28	0.86	0.47	1.00	-0.43	-0.61	-0.13	0.41
LAnkle (Front)	-0.31	-0.34	0.07	-0.33	0.30	-0.43	1.00	0.23	-0.31	0.00
RAnkle (Front)	-0.01	-0.46	-0.08	-0.54	-0.34	-0.61	0.23	1.00	0.13	-0.57
LAnkle (Back)	0.10	0.24	-0.23	0.04	-0.52	-0.13	-0.31	0.13	1.00	0.17
RAnkle (Back)	0.03	0.54	-0.30	0.45	0.08	0.41	0.00	-0.57	0.17	1.00

3.4.2 Distance measurement

Since we proposed two approaches that used time-dependent and time-independent features, then we employ the Dynamic Time Warping (DTW) and Euclidean distance (EU).

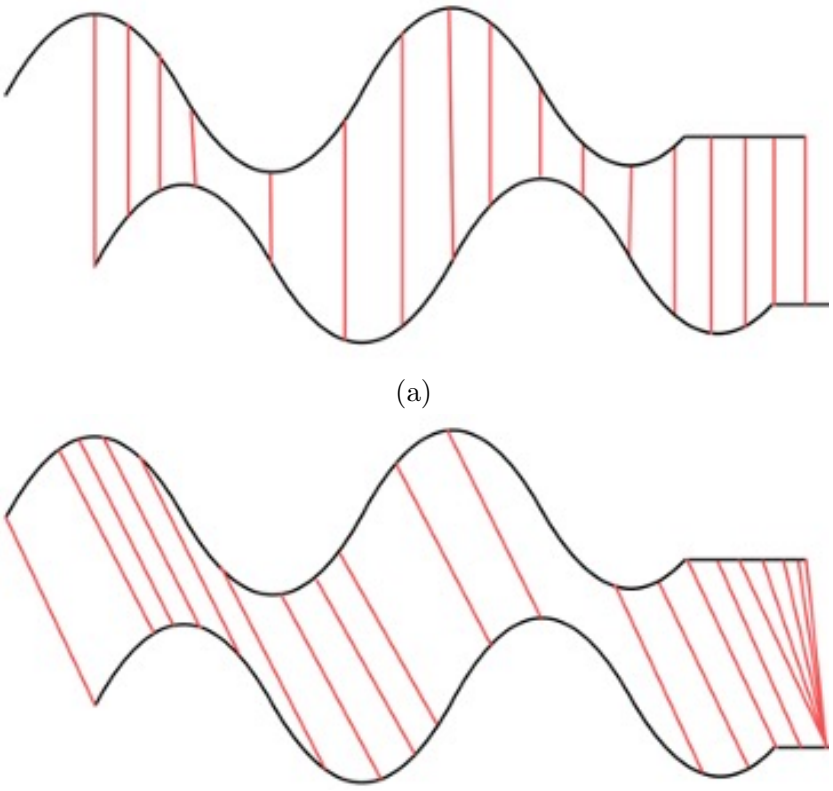
Dynamic Time Warping (DTW) is an algorithm to measure the distance between time series, which can be used to find similarities. This algorithm can handle varying walking speeds and endure time shifts between two sequences. This algorithm is versatile and can be used for different recognition tasks, such as speech and signature recognition, as in the work of C. S. Myers and L. R. Rabiner [37].

DTW offers the most affordable and optimum option for two sequences to be aligned, known as the DTW distance ($S^{i,D}_{DTW}$). Figure 3.11 shows an example of the DTW warping path on the cost matrix of the right hip angle of reference (y-axis) and target sequences (x-axis) between the same person (Figure 3.11a) and a different person (Figure 3.11b).

In fact, the EU is able to match the patterns. However, its straightforward nature makes it unsuitable for analyzing time series data, particularly when comparing sequences that vary in time or speed. This refers to the situation where two corresponding points in sequence do not perfectly line up. Additionally, most of the time series data have different lengths, and Euclidean distance requires sequences to have the same length, making it less suitable for variable-length data. This makes DTW more suitable for use with time series data because it enables time warping to align two patterns, which are walking patterns that vary in time, speed, and sequence

length. Figure 3.10 presents the different patterns of alignment by Euclidean distance (Figure 3.10a) and DTW (Figure 3.10b).

By this reason, the time-independent correlation feature is suitable for EU since the length of two data are fixed to be the same, and the entire sequences are perfectly lined up. In this situation, the measured distance from DTW or EU are the same. However, it has no specific reason for DTW because the time warping characteristic is not required. We conduct this approach to see that weather the time is a requirement for walking pattern matching or else.



(a)

(b)

Figure 3.10: Comparison between Euclidean distance and DTW distance alignments. (a) Direct patterns alignment of Euclidean distance algorithm. (b) Time warping patterns alignment of DTW algorithm.

3.4.3 Matching algorithm

After determining the distance, we match the person identity in a target with reference sequences by finding a minimum distance as in Equation (3.16).

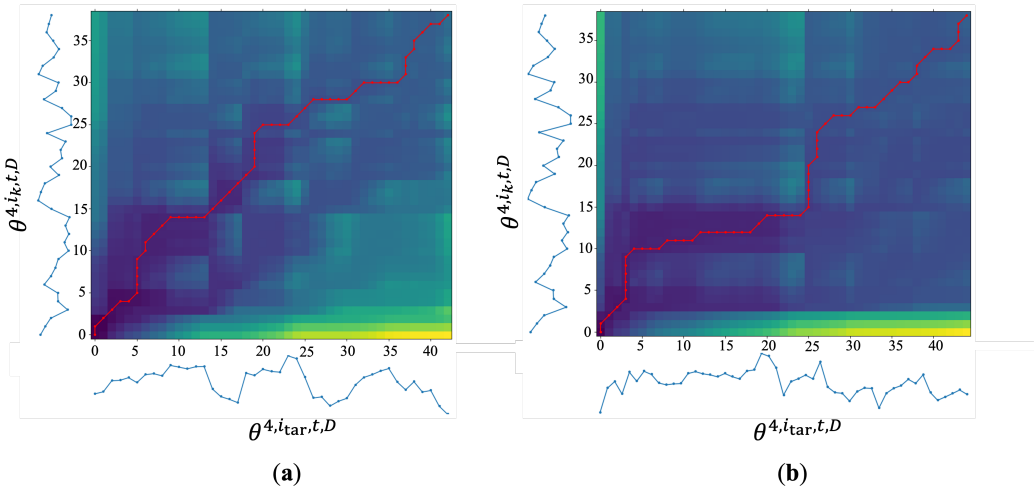


Figure 3.11: Sample of the DTW warping path on the cost matrix of right hip angle at $D = 162°$. (a) DTW warping path with the same person. (b) DTW warping path with a different person.

Since it has multiple cameras for multi-view gait analysis, we let i_k^D represent the matched person identity from each camera perspective (D).

$$i_k^D = \underset{i_k}{\text{argmin}}(S^{i,D}) \tag{3.16}$$

3.4.4 Voting algorithm

Since the multi-view databases use multiple cameras, we obtain multiple matched identities. This implies that the matching accuracy depends on the camera perspective. We then apply majority voting to aggregate the identity from each D by selecting the most frequently appearing identity in every view, as in Figure 3.12a, on the other hand, it is a modal identity. The 'vote' in Equation (3.17) refers to the mentioned voting algorithm. In fact, it is simply a mode in statistics [18].

$$i_k = \text{vote}\{i_k^{0°}, ..., i_k^D\} \tag{3.17}$$

In the case that it have no modal identity, i_k will be selected from the main camera that achieves highest accuracy, as an example shown in Figure 3.12b. It is a situation when modal identity is unavailable, then it select the i_k^D from stared camera to be i_k.

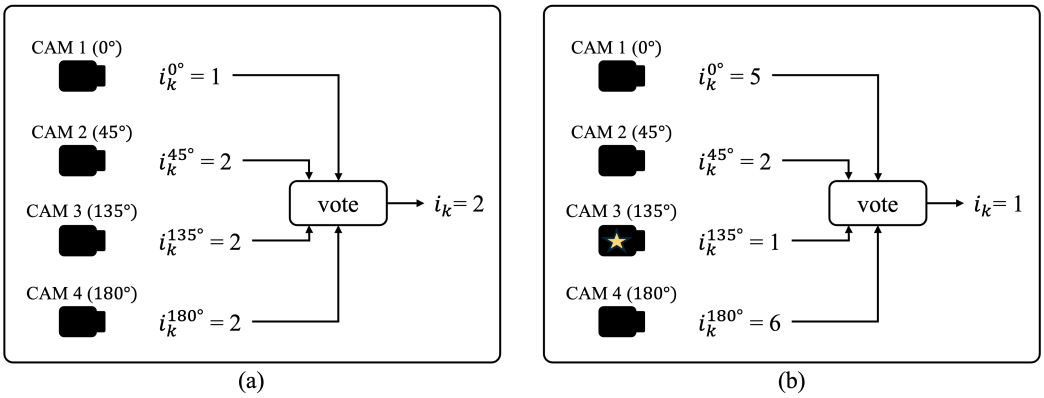


Figure 3.12: Example of the voting situation to describe the procedure to obtain i_k by 'vote'. (a) Example of the case where modal identity is available. (b) Example of the case where modal identity is unavailable.

3.4.5 Accuracy measurement

We evaluate the proposed method by calculating the accuracy in equations (3.18) and (3.19) to measure the correctness of the matched identities. Equation (3.18) is used to calculate the accuracy of the matched identities without a majority vote, and equation (3.19) is used when applying a majority vote.

$$\text{Accuracy (without voting)} = \frac{\Sigma i_{k,c}^{D}}{|\mathbb{K}|} \tag{3.18}$$

$$\text{Accuracy (with voting)} = \frac{\Sigma i_{k,c}}{|\mathbb{K}|} \tag{3.19}$$

Chapter 4

Experiments and results

This chapter includes all experiment results. This chapter will begin by introducing the objectives, experiment methods, and datasets used in the experiments. Then, followed the results obtained from these experiments.

The objectives of this experiment are defined as follows:

- To analyze human motion from the multi-view gait image.
- To improve the human gait analysis method from view-variation problem.
- To study the effect of the human body part features to an analysis of the gait.
- To find an importance of each joint feature to an analysis of the gait.
- To find the optimal approach for identification.

In this study, we divide the experiments into 6 parts as follows:

1. The significance of different body parts determination.
2. Robustness of the different body parts features.
3. The significance of different joints determination.
4. Comparative results of different voting algorithms.
5. Comparative results between distance measurement algorithms.
6. Comparative results with prior studies.

4.1 Experimental conditions

4.1.1 Datasets

- *CASIA-B dataset*

We apply our method to the CASIA-B dataset [55]. It is a multi-view gait database that captures 124 subjects. This study implements the sequences of $—\mathbb{K}— = 118$ subjects. It includes $—\mathbb{D}— = 11$ cameras that equipped at eye-level. Its perspectives (D) spans from 0° to 180° with 18° intervals, as shown in Figure 4.1a.

All sequences captured the subjects walking from a starting point to the marked endpoint with cameras equipped at eye level, as shown in Figure 3.6a. The 0° captured a frontal perspective, the 90° captured a side perspective, and the 180° captured a rear perspective. It consists of three walking conditions, i.e., normal walking (NM), walking while carrying a shoulder bag (BG), and walking while wearing a down coat (CL), as shown in Figures 4.1b–4.1d. In this study, our interest is the NM condition.

There are six sub-datasets containing in the NM condition as NM01–NM06. We employ the NM01 to be a reference sequences and NM02 to be a target sequences.

- *OUMVLP-Pose dataset*

The OUMVLP-Pose is an OU-ISIR gait database with extracted 2D pose estimation ($p_i^j(x, y, t)$) by OpenPose and AlphaPose [2]. It contains sequences of 10,307 subjects walking round trip. This study employ $\mathbb{K} = 100$ subjects that captured by 14 cameras, which D spanning from 0° to 270° with 15° intervals as shown in Figure 4.2. The cameras are equipped at a higher position, making the captured images of OUMVLP-Pose similar to the real surveillance scenarios.

For this dataset, we employ two sub-datasets (OP01 and OP02) that extract the joints using OpenPose. We use one sub-dataset (OP01) as a reference sequence and another as a target sequence (OP02). For the sub-dataset that used AlphaPose to estimate human joints, we employed AP01 as reference sequences and AP02 as target sequences for matching. To perform pattern matching, we use two sub-datasets for experimentation, one as a reference and one as a target. Even if there are more than two sub-datasets, we select only two of them.

It is crucial to note that we match the reference and target sequences under the identical view. The reason is that the proposed method is pattern matching, and it has no feature learning state. Basically, accuracy will

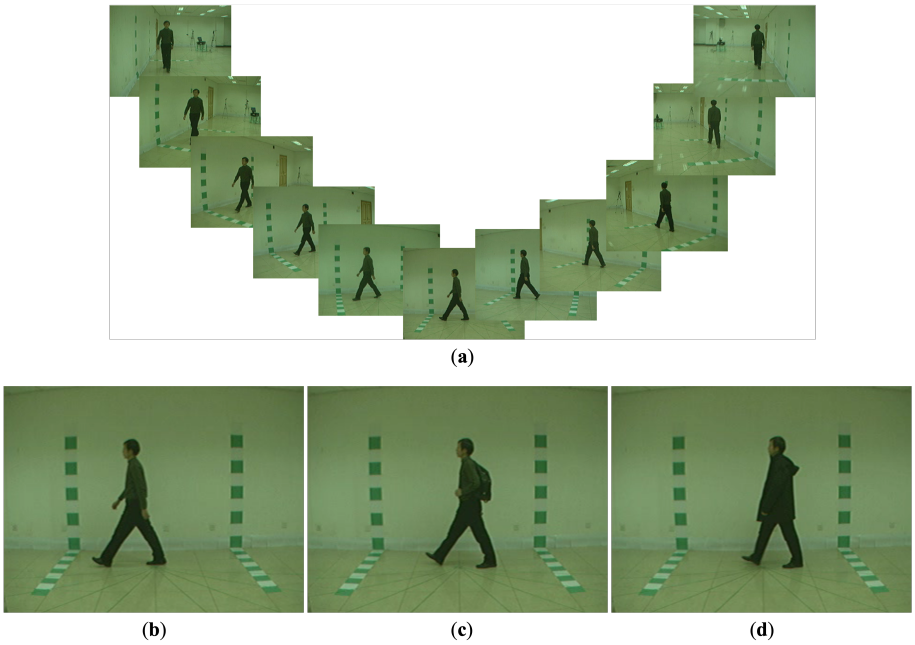


Figure 4.1: Samples of a multi-view CASIA-B gait database [55]. (a) Gait images from the different camera perspectives. (b) Normal walking condition (NM sub-dataset). (c) Walking with carrying condition (BG sub-dataset). (d) Walking with clothing condition (CL sub-dataset)

decrease when compared across different perspectives. Therefore, our main focus is to overcome the view-variation by integrating the multiple perspectives using voting method.

4.1.2 Parameters adjustment

- *Eye-level scenario*

We let the NM sub-dataset from CASIA-B multi-view gait database to be a representative of eye-level scenario in this experiment. It is a scenario where cameras are equipped at the same level of human's eyes. In fact, this scenario is enriched of the gait information, but not practical enough for the real situation application.

The number of subjects $\mathbb{K}$ are divided into three conditions, i.e., 20, 49, and 118 subjects.

- *Surveillance scenario*

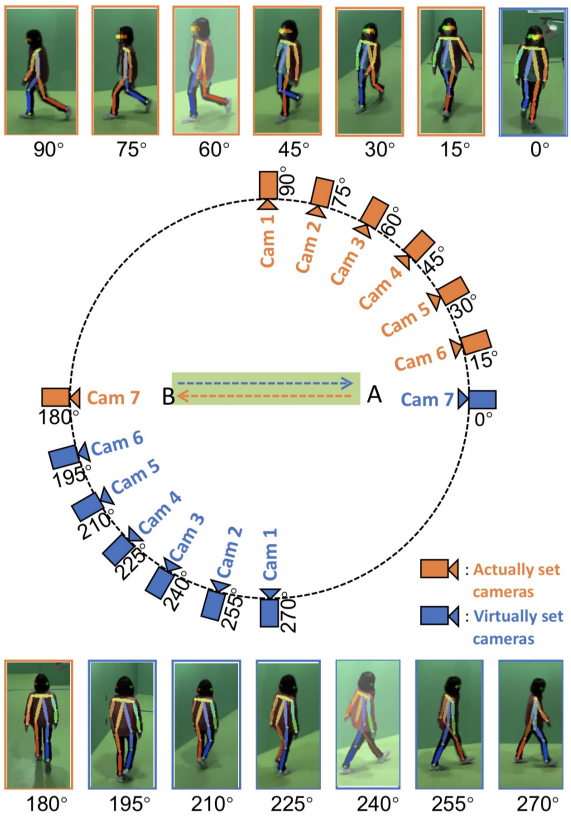


Figure 4.2: Capturing setup environment of OUMVLP-Pose [2] and samples images with extracted human pose estimation. The "actually set cameras" implies to the cameras perspectives that capture subjects walking from A to B, and vice versa on "virtual set cameras".

The OUMVLP-Pose dataset takes a representation of surveillance scenario. Since the cameras are equipped at a higher level, which is the same perspective as real surveillance cameras. Generally, this scenario brings more challenges to the research due to self occlusion issues. However, its perspective is practical in real-world situation.

We divided $|\mathbb{K}|$ into 20, 50, and 100 subjects from the OUMVLP-Pose dataset.

4.1.3 Pose estimation algorithm

- *Eye-level scenario*

Since the CASIA-B dataset is provided only the video sequences, MediaPipe is selected to be a pose estimation algorithm to extract the 2D joints coor-

dinate in x-axis and y-axis, and 3D joints coordinate in x-axis, y-axis, and z-axis.

We can extract ten angles by using MediaPipe as a pose estimation algorithm including elbow, hip, knee, front and back ankle, both left and right sides. The number of in Equation (3.9).

- *Surveillance scenario*

The OUMVLP-Pose dataset already provides 18 joint landmarks as 2D joint coordinates in x-axis and y-axis from OpenPose and AlphaPose. Unfortunately, it lack of the video sequences, makes it impossible to apply MediaPipe with this dataset.

Furthermore, the human landmark estimation model does not allow for the extraction of ankle angles. Thus, the number of joint angles of this dataset is $|\mathbb{U}| = 6$ angles.

4.2 The significance of different body parts determination

In this experiment, we separate the features into three parts with respect to the body parts, i.e., whole, upper, and lower body, to study the significance of the body parts to an analysis of the gait in each scenario.

We propose two approaches that use DTW and EU to match patterns based on time-dependent and time-independent features, respectively. Notably, the results without majority vote refer to the results of the matching on a single perspective, and with a majority vote is a result after applying majority vote for views integration.

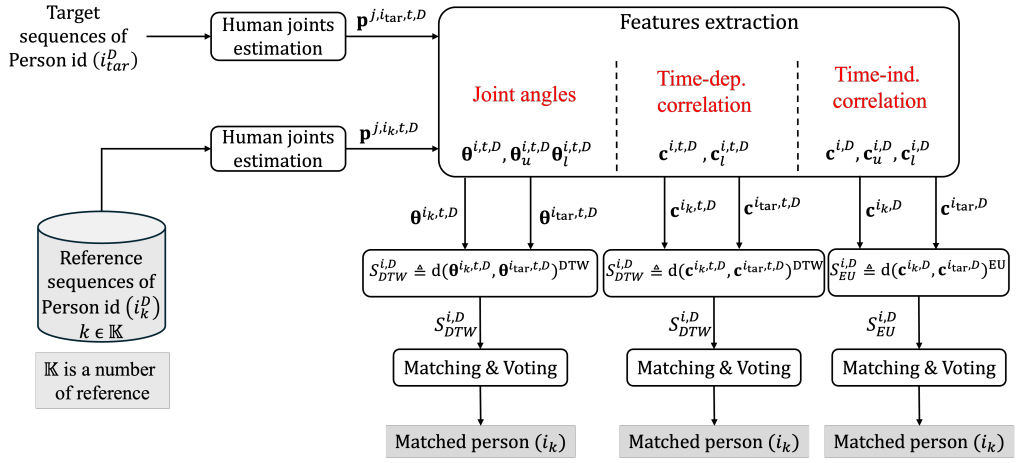


Figure 4.3: Diagram to describe the process of 3 approaches in the significance of different body parts determination experiment.

4.2.1 Approach 1 & 2: Apply Dynamic Time Warping (DTW) with time-dependent features.

This approach employs DTW with time-dependent features to match the patterns. These features include joint angles (Approach 1) and time-dependent correlation (Approach 2). It will be separated into three vectors with respect to the body parts, as shown in Figure 4.3.

- *Eye-level scenario*

Figure 4.4a presents the accuracy of the matching without majority vote, which Figures 4.4a–4.4c display the accuracy without majority vote when

employing the joint angles as a feature of $\mathbb{K} = 20$, 49, and 118 subjects, and Figures 4.4d–4.4f show the accuracy without majority vote when employing the time-dependent correlation as a feature of $\mathbb{K} = 20$, 49, and 118 subjects, respectively. Notably, the entire results are shown in tables 5.1-5.2 in Appendix.

Figure 4.5 presents the accuracy with majority vote when employing joint angles as a feature (Figure 4.5a), and time-dependent correlation as a feature (Figure 4.5b).

In this scenario, we found that the accuracy of whole body and upper body parts increases and becomes stable when $D \geq 54°$, as shown in Figure 4.4a–4.4c. It indicates that gait information is enriched from these perspectives. Moreover, increasing the number of $\mathbb{K}$ affects the decreasing accuracy. Normally, noise or irrelevant data will be normalized when the number of samples is increased. However, it might suppress individual data by treating it as noise, and it reduces the accuracy of the matching.

Figure 4.5a shows that the upper body part achieves the highest accuracy, but it is far enough to conclude that the upper part is the most important part in gait recognition since it includes only two angles. Generally, it is impossible to identify people by observing only their arm swings. This could be due to an outlier or data overfitting.

Figures 4.4d–4.4f and Figure 4.5b indicate that the time-dependent correlation fails to identify the person's identity. This is because the data points are not varied enough to represent the relationships and associations between them. The upper body, consisting of only two angles, is incapable of determining the time-dependent correlation between two data points. Furthermore, when combining both features, i.e., the joint angles and time-dependent correlation, it is unable to improve the performance and accuracy of the matching, as shown in Figure 4.6. According to the time-dependent correlation itself, it fails to identify people, making it useless when combined with another feature. The entire result is shown in table 5.3 in Appendix.

Additionally, we found that the accuracy of the matching is increased when employing the 3D joints extracted from MediaPipe (Figure 4.7b). Compared to 2D joints, the accuracy of 118 subjects is increased from 48% to 66% (increased relatively by 37.5%). For 3D joint angles, the whole body part is the most reliable features as same as 2D joints. The entire result for 3D feature is presented in table 5.5 in Appendix.

In this case, the whole body part of the joint angles feature is significant for identifying identities based on the walking pattern matching.

- *Surveillance scenario*

This scenario is critical, but it presents the greatest challenge to gait analysis because it is similar to the real-world situation where cameras are mostly equipped in higher positions (over the head of a human). This scenario holds significant importance in the study of human gait patterns. In this experiment, the OUMVLP-Pose dataset will be representative of this scenario.

Figures 4.8a–4.8f show the accuracy without majority vote on the OUMVLP-Pose dataset that joints extracted by OpenPose. Figures 4.8a–4.8c present the accuracy without majority vote when using joint angles as a feature, and Figures 4.8d–4.8f present the accuracy without majority vote when using time-dependent correlation as a feature of 20, 50, and 100 subjects.

Figures 4.8a–4.8c suggest that the joint angles of the lower body achieve the most reliable and consistent part. It might be because in this scenario, the lower body parts are less occluded by the upper body parts, which makes them more clearly visible, unlike in the eye-level scenario. Unfortunately, Figures 4.8d–4.8f indicate that approach 2 is unable to identify person's identity.

The results, with a majority vote that shown in Figure 4.9a, imply that the lower body feature of joint angles serves better in pattern matching. In addition, it is less variable in the number of $\mathbb{K}$ compared with the whole and upper body features. Since the whole body feature includes both upper and lower body parts, but the upper body is not significant to improve accuracy and is inconsistent, the whole body feature becomes unreliable and drastically affected by the number of $\mathbb{K}$, similar to the upper body feature. In this case, the lower body joint angles feature is the most significant feature for identifying a person's identity based on walking pattern matching.

Unfortunately, the time-dependent correlation also fails to identify people, especially when using OpenPose (in Figure 4.9b), for the same reason described in the eye-level scenario. Figure 4.10 presents the results after combining both joint angles and time-dependent correlation together. However, it is not significant enough to improve the match.

Additionally, Figures 4.11a–4.11f show the accuracy without a majority vote in which joints were extracted by AlphaPose. Figures 4.11a–4.11c present the accuracy without majority vote when using joint angles as a feature, and Figures 4.11d–4.11f present the accuracy without majority vote when using time-dependent correlation as a feature of 20, 50, and 100 subjects, respectively. Figures 4.12a and 4.12b display the accuracy with a majority vote based on the extracted joints from AlphaPose when using joint angles and time-dependent correlation as a feature, respectively.

For this case, the whole body is the most significant part, unlike the OpenPose case. Similarly, approach 2 fails to identify person's identity, even though we combine it with joint angles, it still unable to use (see the results in Figure 4.13). The results for OUMVLP-Pose can be found in tables 5.7

and 5.14

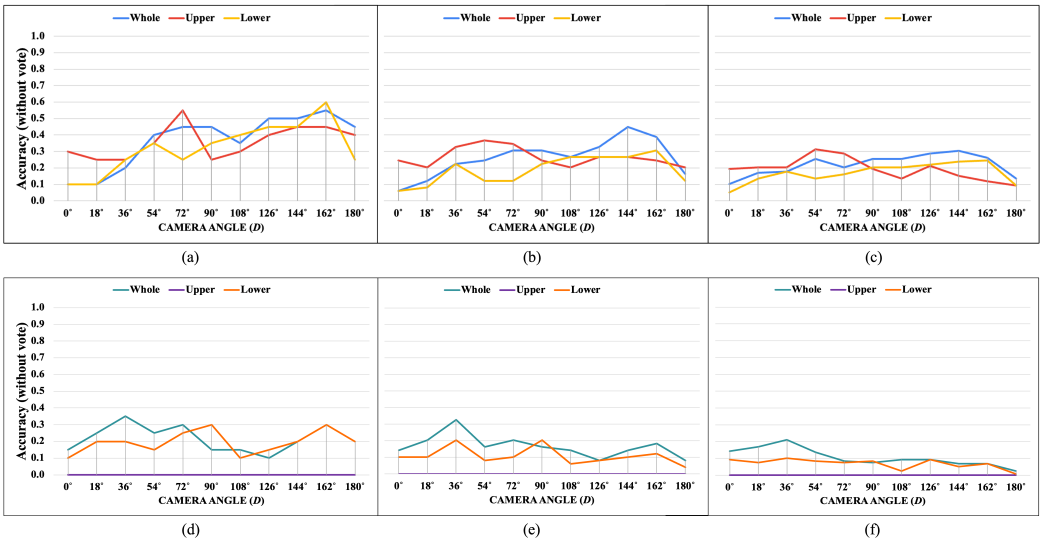
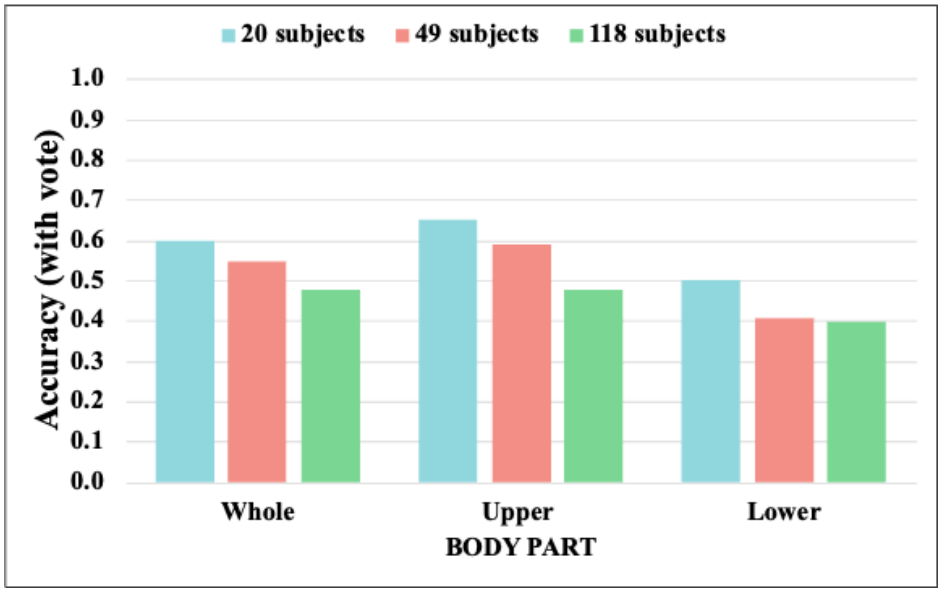
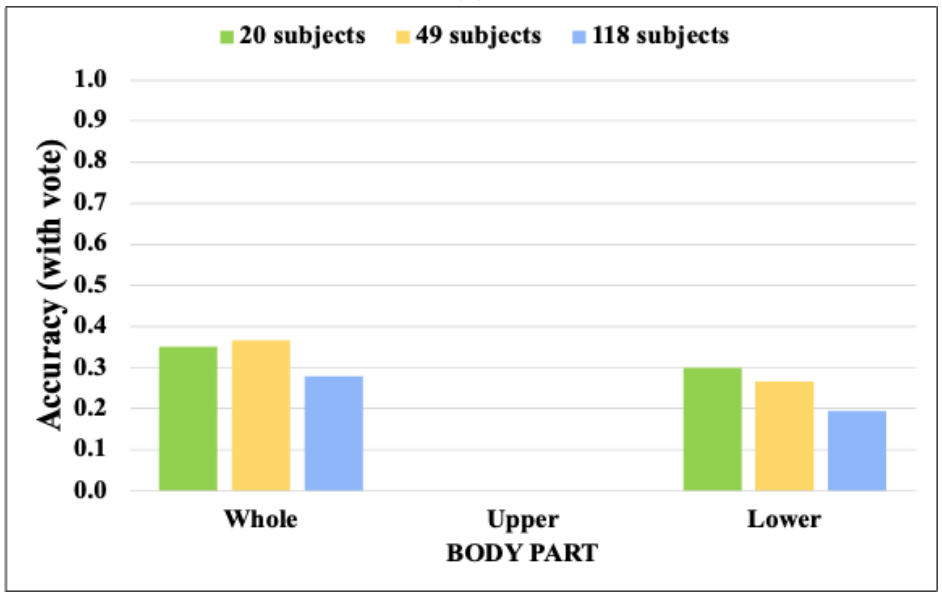


Figure 4.4: Accuracy of the matching without majority vote on NM sub-dataset. These results are from employing 2D joints extracted by MediaPipe. (a) Accuracy of the joint angles being used as a feature of 20 subjects. (b) Accuracy of the joint angles being used as a feature of 49 subjects. (c) Accuracy of the joint angles being used as a feature of 118 subjects. (d) Accuracy of the time-dependent correlation being used as a feature of 20 subjects. (e) Accuracy of the time-dependent correlation being used as a feature of 49 subjects. (f) Accuracy of the time-dependent correlation being used as a feature of 118 subjects.



(a)



(b)

Figure 4.5: Accuracy of the matching with majority vote on NM sub-dataset. These results are from employing 2D joints extracted by MediaPipe. (a) Accuracy with majority vote of the joint angles being used as a feature of 20, 49 and 118 subjects. (b) Accuracy with majority vote of the time-dependent correlation being used as a feature of 20, 49 and 118 subjects.

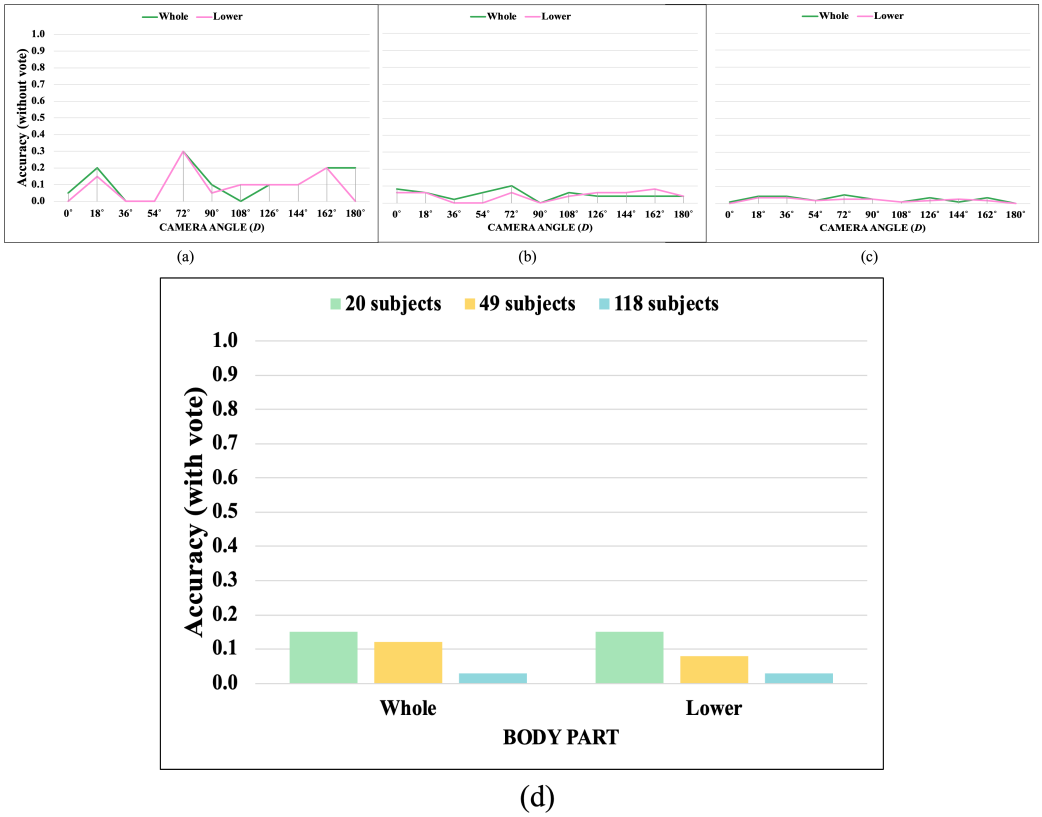


Figure 4.6: Accuracy of the matching with majority vote on NM sub-dataset. These results are from employing 2D joints extracted by MediaPipe. (a) Accuracy with majority vote of the joint angles and time-dependent correlation features being used of 20 subjects (b) Accuracy with majority vote of the joint angles and time-dependent correlation features being used of 49 subjects (c) Accuracy with majority vote of the joint angles and time-dependent correlation features being used of 118 subjects. (d) Accuracy with majority vote of 20, 49 and 118 subjects.

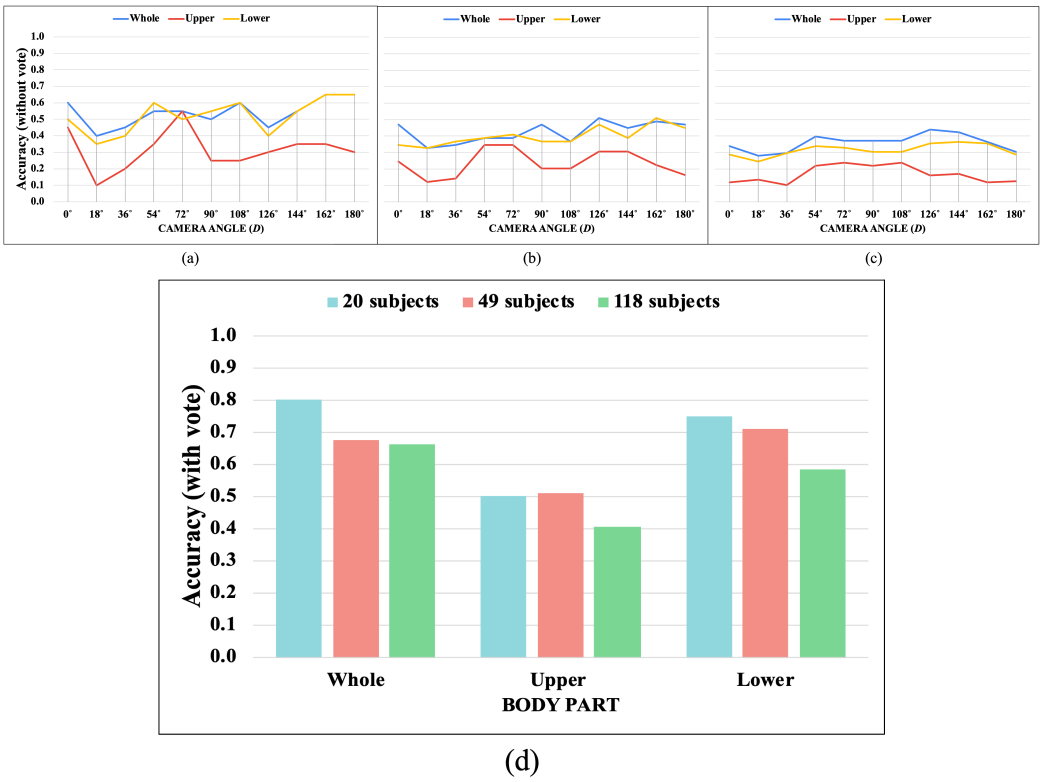


Figure 4.7: Accuracy of the matching with majority vote on NM sub-dataset. These results are from employing 3D joints extracted by MediaPipe. (a) Accuracy without majority vote of the joint angles being used as a feature of 20 subjects (b) Accuracy without majority vote of the joint angles being used as a feature of 49 subjects (c) Accuracy without majority vote of the joint angles being used as a feature of 118 subjects. (d) Accuracy with majority vote of 20, 49 and 118 subjects.

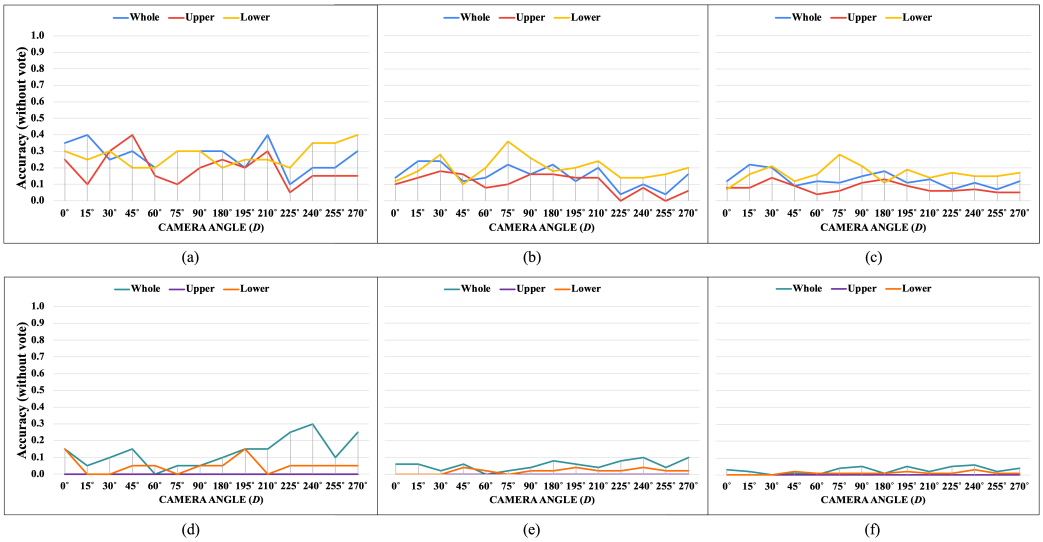
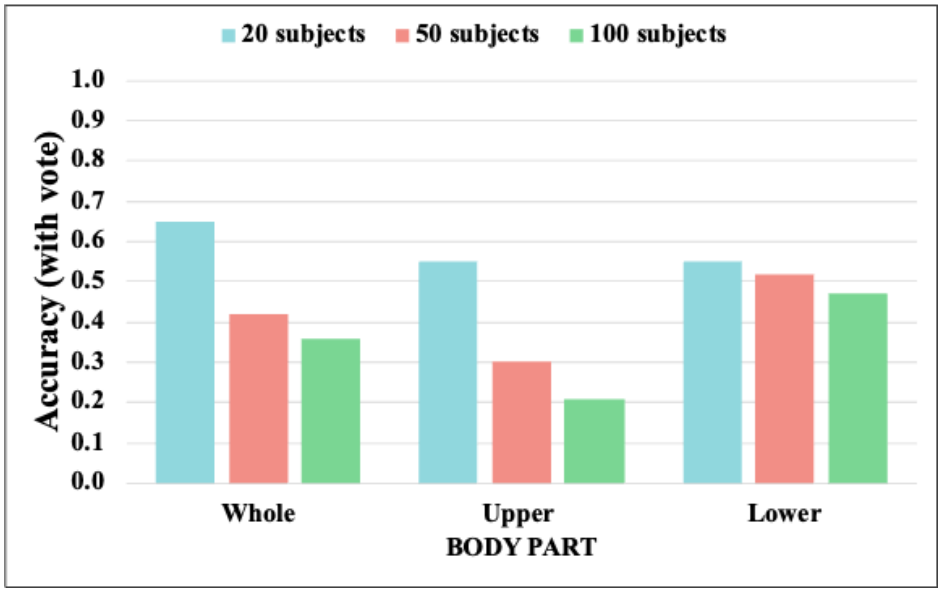
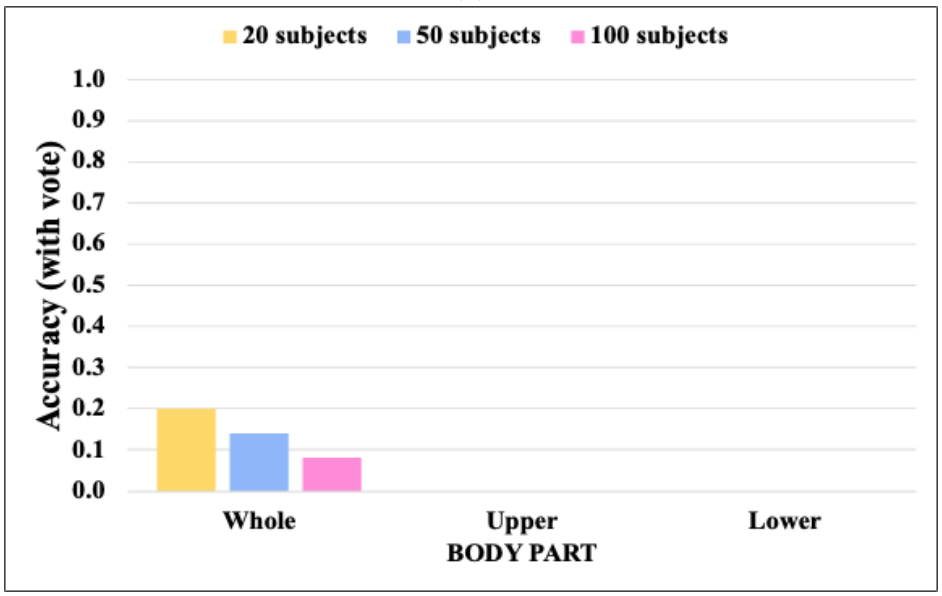


Figure 4.8: Accuracy of the matching without majority vote on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by OpenPose. (a) Accuracy of the joint angles being used as a feature of 20 subjects. (b) Accuracy of the joint angles being used as a feature of 50 subjects. (c) Accuracy of the joint angles being used as a feature of 100 subjects. (d) Accuracy of the time-dependent correlation being used as a feature of 20 subjects. (e) Accuracy of the time-dependent correlation being used as a feature of 50 subjects. (f) Accuracy of the time-dependent correlation being used as a feature of 100 subjects.



(a)



(b)

Figure 4.9: Accuracy of the matching with majority vote on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by OpenPose. (a) Accuracy with majority vote of the joint angles being used as a feature of 20, 50 and 100 subjects. (b) Accuracy with majority vote of the time-dependent correlation being used as a feature of 20, 50 and 100 subjects.

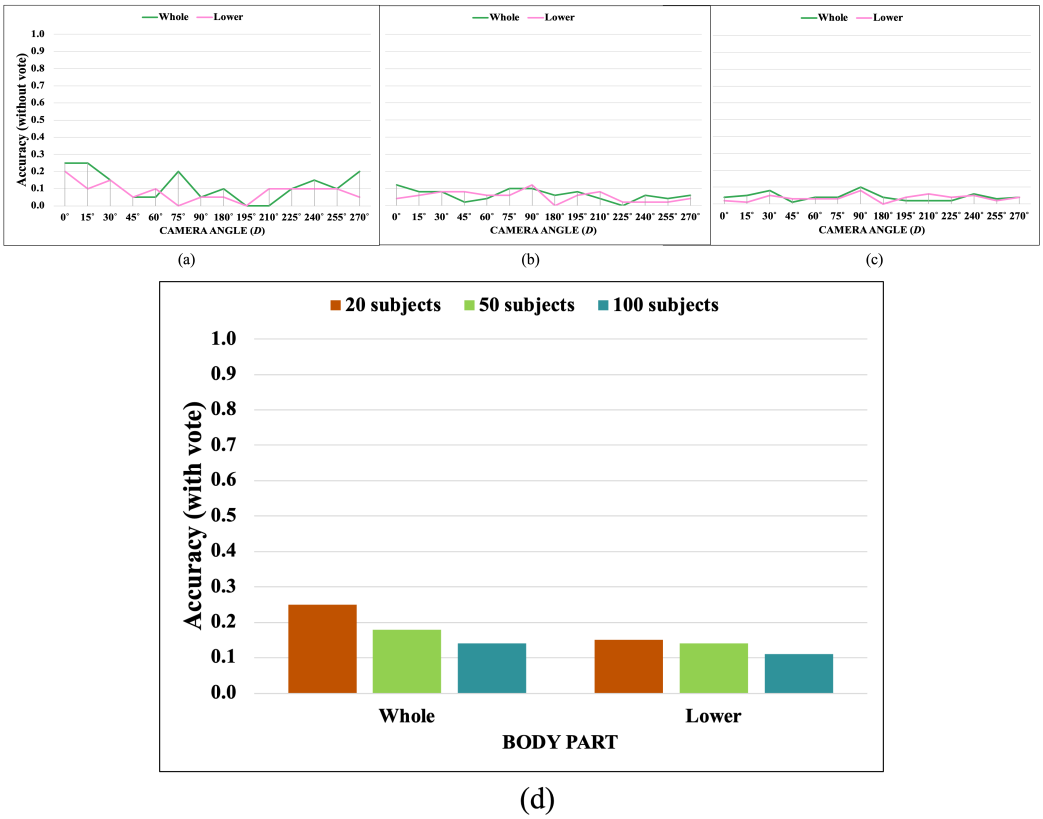


Figure 4.10: Accuracy of the matching with majority vote on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by OpenPose. (a) Accuracy with majority vote of the joint angles and time-dependent correlation features being used of 20 subjects (b) Accuracy with majority vote of the joint angles and time-dependent correlation features being used of 50 subjects (c) Accuracy with majority vote of the joint angles and time-dependent correlation features being used of 100 subjects. (d) Accuracy with majority vote of 20, 50 and 100 subjects.

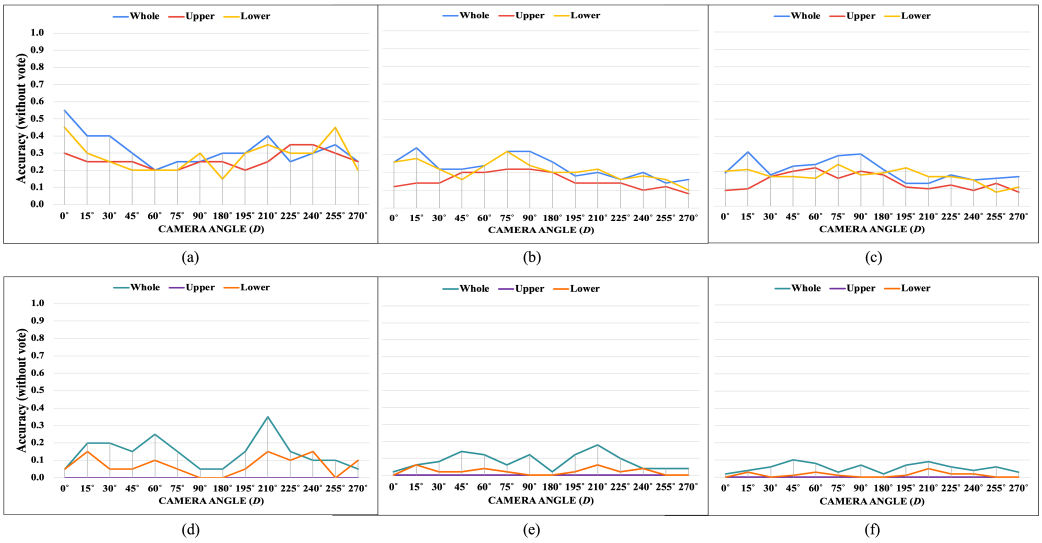
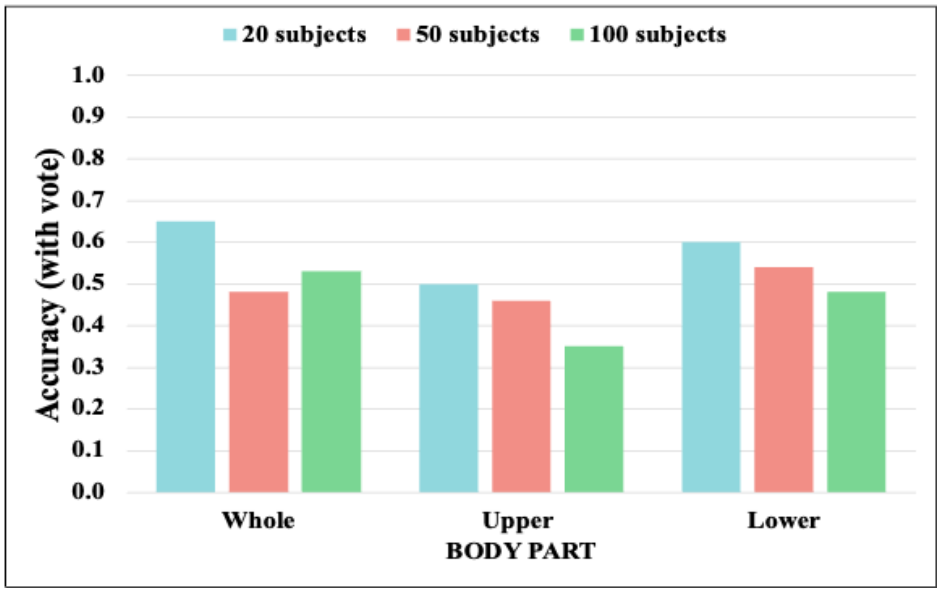
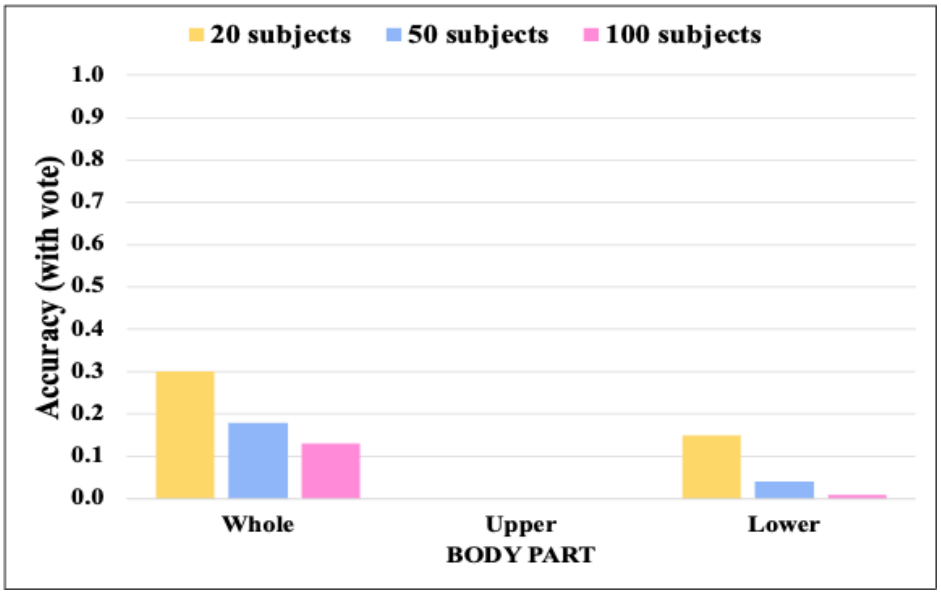


Figure 4.11: Accuracy of the matching without majority vote on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by AlphaPose. (a) Accuracy of the joint angles being used as a feature of 20 subjects. (b) Accuracy of the joint angles being used as a feature of 50 subjects. (c) Accuracy of the joint angles being used as a feature of 100 subjects. (d) Accuracy of the time-dependent correlation being used as a feature of 20 subjects. (e) Accuracy of the time-dependent correlation being used as a feature of 50 subjects. (f) Accuracy of the time-dependent correlation being used as a feature of 100 subjects.



(a)



(b)

Figure 4.12: Accuracy of the matching with majority vote on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by AlphaPose. (a) Accuracy with majority vote of the joint angles being used as a feature of 20, 50 and 100 subjects. (b) Accuracy with majority vote of the time-dependent correlation being used as a feature of 20, 50 and 100 subjects.

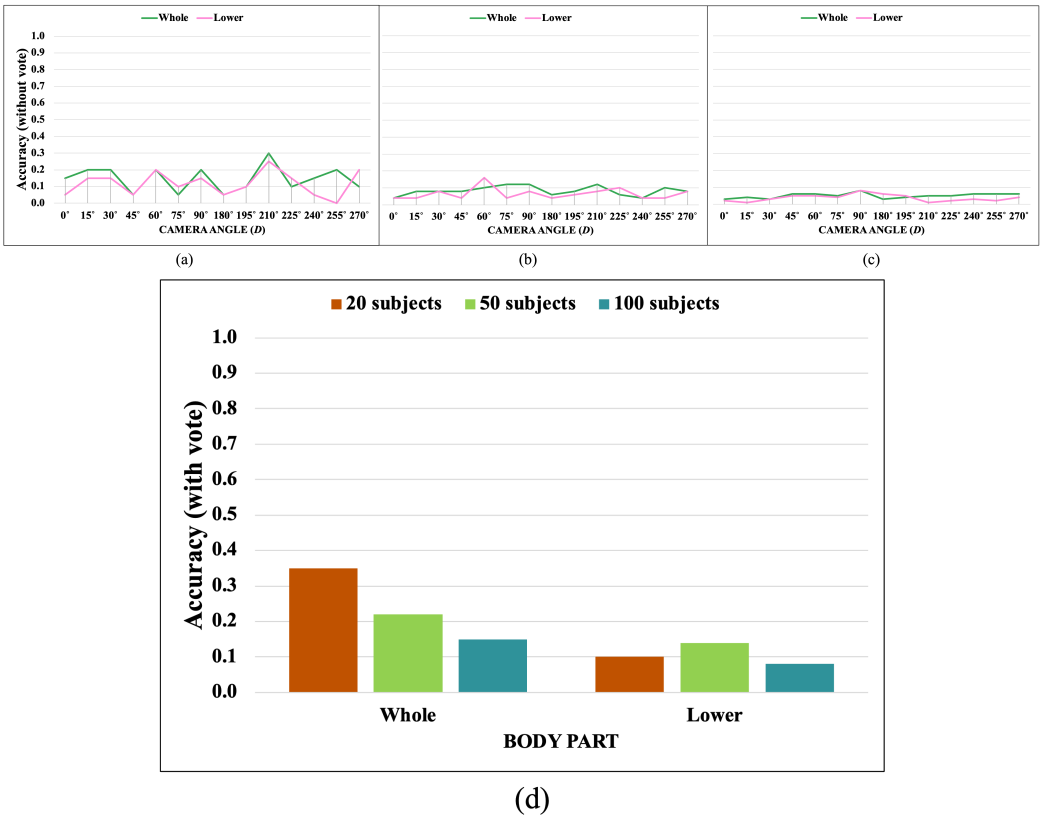


Figure 4.13: Accuracy of the matching with majority vote on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by AlphaPose. (a) Accuracy with majority vote of the joint angles and time-dependent correlation features being used of 20 subjects (b) Accuracy with majority vote of the joint angles and time-dependent correlation features being used of 50 subjects (c) Accuracy with majority vote of the joint angles and time-dependent correlation features being used of 100 subjects. (d) Accuracy with majority vote of 20, 50 and 100 subjects.

4.2.2 Approach 3: Apply Euclidean distance (EU) with time-independent feature.

This approach uses a time-independent feature, eliminating the need for time-warping during matching. As a result, we use direct matching (EU) to match the patterns. The only feature for this approach is a time-independent correlation, which will be separated into three vectors with respect to the body parts as shown in Figure 4.3.

- *Eye-level scenario*

Figure 4.14 presents the accuracy with and without majority vote on NM sub-dataset of CASIA-B. Figures 4.14a–4.14c show the accuracy without majority vote, and Figure 4.14d show the accuracy with majority vote.

The whole body feature consistently outperforms other features in terms of accuracy, both with and without voting. Unfortunately, the upper body feature proves ineffective in identifying individuals. This is because, despite an increase in the number of data points, the upper body still only includes two angles. It indicates that even if we increase the number of data points in each variable to make it possible for correlation calculation, the results prove that to represent a walking pattern, it requires as many variables (or joint angles). This also answers the question of why the whole body is the most significant part, even if the upper body fails, unlike in approach 1.

However, time-independent correlation is still affected by the variations of $\mathbb{K}$, and the accuracy is reduced when the number of $\mathbb{K}$ is increased. Furthermore, the accuracy differs depending on the perspective. Even though it seems like a reliable feature, it appears to suffer the most with camera perspectives.

When employing the 3D joints for calculating the time-independent correlation, we found that the accuracy slightly decreased from 74% to 67% (it decreased by 9.46%) on 118 subjects (see Figure 4.15d). This might be because of the error in joints estimation, especially the z-axis. Since MediaPipe uses DNN to estimate the z-axis of the human joints on a single image from a single perspective, it requires at least stereo vision for maximum reliability. On top of that, the whole body of the 3D feature achieves the highest accuracy, yet is more consistent than the lower body feature.

To conclude this, the whole body feature of time-independent correlation is the most significant feature in this scenario.

- *Surveillance scenario*

Figure 4.16 presents the accuracy without and with majority vote on the OUMVLP-Pose dataset, which applies OpenPose for joints extraction. Figures 4.16a–4.16c show the accuracy without a majority vote, and Figure 4.16d show the result with a majority vote.

Figure 4.17 shows the accuracy without and with a majority vote on the OUMVLP-Pose dataset when applying the AlphaPose to extract the joints. Figures 4.17a–4.17c show the accuracy without a majority vote, and Figure 4.17d show the result with a majority vote.

In this scenario, the whole body is the best part of the feature for walking pattern matching, just as in the eye-level scenario, for the same reasons described in the previous section.

However, its weakness is that it is more affected by variations in camera perspectives. Moreover, the result with a majority vote suggests that the number of $\mathbb{K}$ dramatically affects the accuracy, as the accuracy drastically decreased when the number of $\mathbb{K}$ increased. It implies that by using time-independent correlation as a feature representative of walking pattern, it loses fine detail that is meaningful for distinguishing people. Furthermore, it suggests that this feature suffers more from outliers and data overfitting than using joint angles directly, especially in this scenario, which is the most challenging for gait analysis.

Let us conclude that the whole body part of the time-independent correlation is the most significant feature in this scenario.

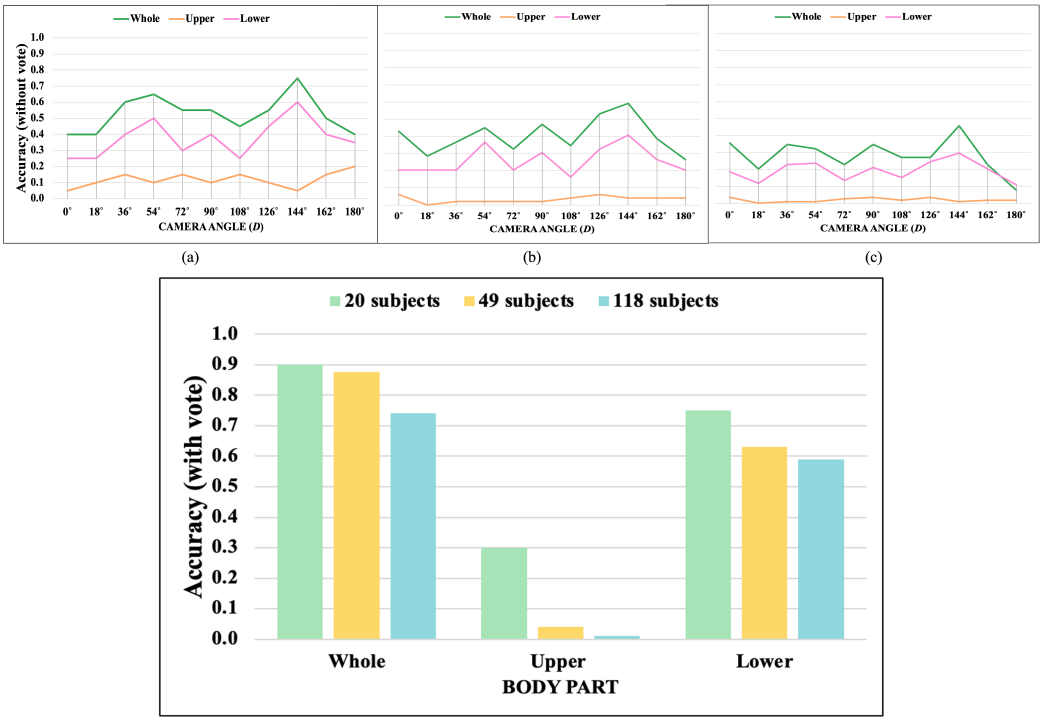


Figure 4.14: Accuracy of the matching without and with majority vote on NM sub-dataset. These results are from employed time-independent correlation based on 2D joints extracted by MediaPipe. (a) Accuracy without voting of the time-independent correlation being used as a feature of 20 subjects. (b) Accuracy without voting of the time-independent correlation as a feature of 49 subjects. (c) Accuracy without voting of the time-independent correlation being used as a feature of 118 subjects. (d) Accuracy with voting of the time-independent correlation being used as a feature of 20, 49, and 118 subjects.

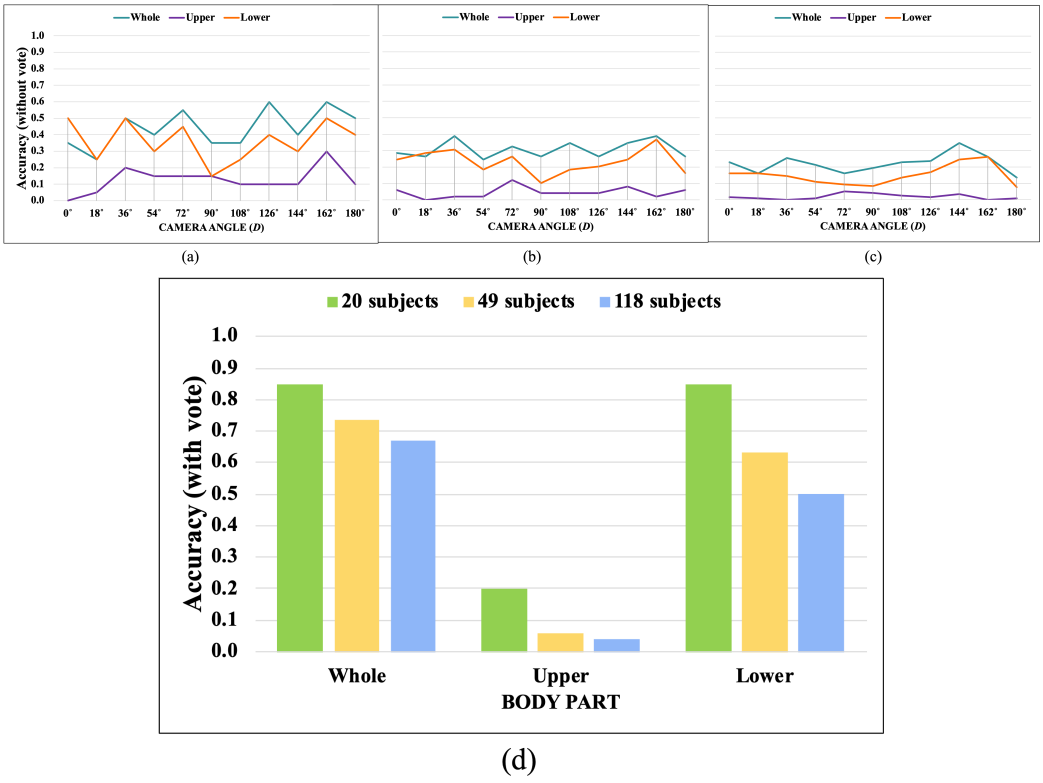


Figure 4.15: Accuracy of the matching without and with majority vote on NM sub-dataset. These results are from employed time-independent correlation based on 3D joints extracted by MediaPipe. (a) Accuracy without voting of the time-independent correlation being used as a feature of 20 subjects. (b) Accuracy without voting of the time-independent correlation as a feature of 49 subjects. (c) Accuracy without voting of the time-independent correlation being used as a feature of 118 subjects. (d) Accuracy with voting of the time-independent correlation being used as a feature of 20, 49, and 118 subjects.

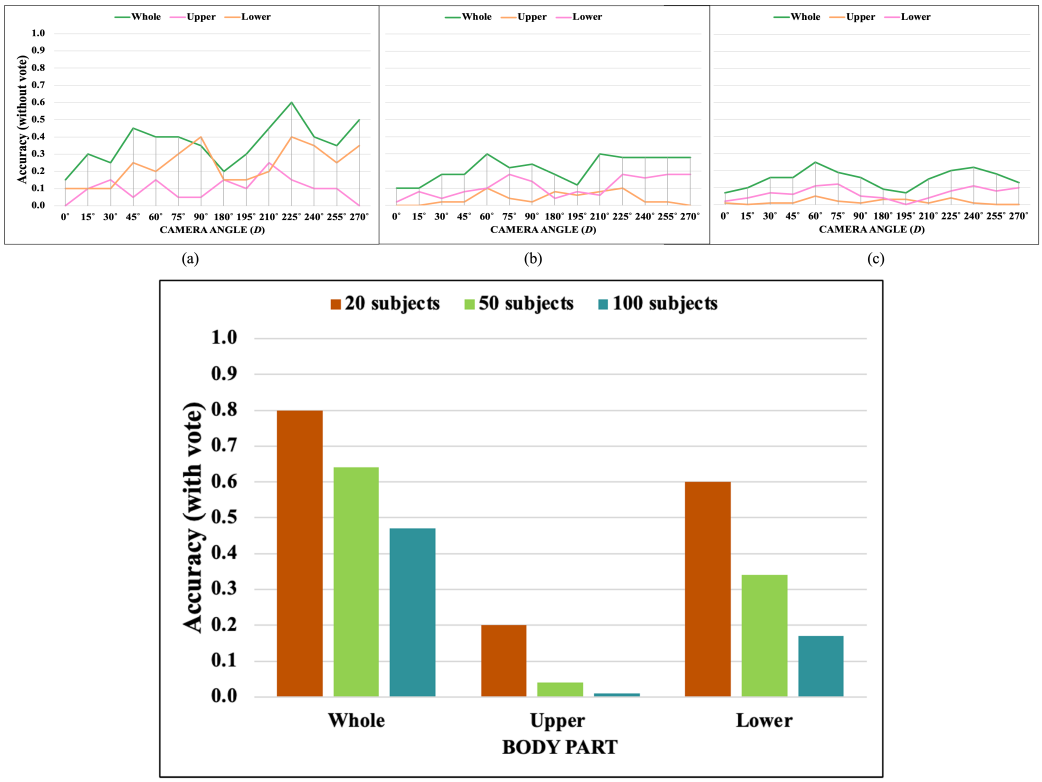


Figure 4.16: Accuracy of the matching without and with majority vote on OUMVLP-Pose dataset. These results are from employed These results are from employed time-independent correlation based on 2D joints extracted by OpenPose. (a) Accuracy without voting of the time-independent correlation being used as a feature of 20 subjects. (b) Accuracy without voting of the time-independent correlation as a feature of 50 subjects. (c) Accuracy without voting of the time-independent correlation being used as a feature of 100 subjects. (d) Accuracy with voting of the time-independent correlation being used as a feature of 20, 50, and 100 subjects.

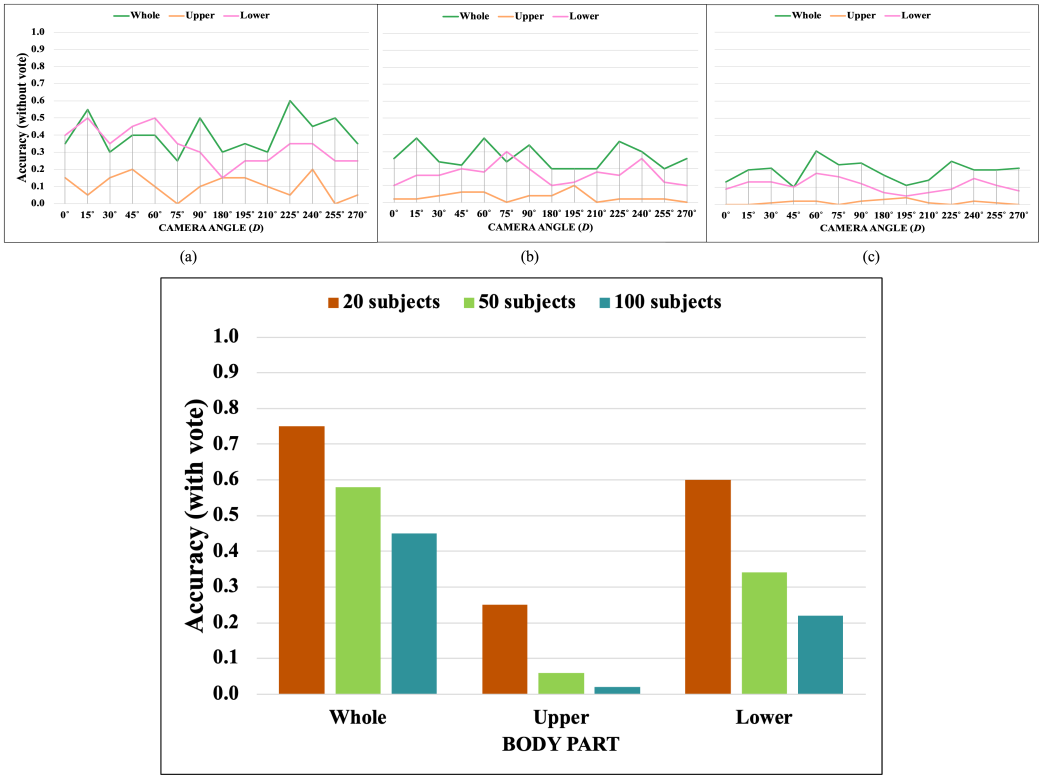


Figure 4.17: Accuracy of the matching without and with majority vote on OUMVLP-Pose dataset. These results are from employed These results are from employed time-independent correlation based on 2D joints extracted by AlphaPose. (a) Accuracy without voting of the time-independent correlation being used as a feature of 20 subjects. (b) Accuracy without voting of the time-independent correlation as a feature of 50 subjects. (c) Accuracy without voting of the time-independent correlation being used as a feature of 100 subjects. (d) Accuracy with voting of the time-independent correlation being used as a feature of 20, 50, and 100 subjects.

4.3 Robustness of the different body parts features

In this experiment, we reduced the number of camera perspectives ($\mathbb{D}$) to half in order to analyze the robustness of the body part features. For the eye-level scenario, we reduce the number of $\mathbb{D}$ from 11 to 5 and 3, which are 0°, 36°, 90°, 144°, and 162°, and 0°, 54°, and 144°, respectively. We also reduce the number of $\mathbb{D}$ from 14 to 7 and 5 for the surveillance scenario, where 15°, 30°, 60°, 90°, 180°, 225°, and 255°, and 15°, 30°, 75°, 180°, and 225° are selected.

This section provides the discussion from employing the whole body feature based on the approach 1 and approach 3, because approach 2 fails to identify person.

4.3.1 Eye-level scenario

In this scenario, we reduce $|\mathbb{D}|$ from 11 to 5 (0°, 36°, 90°, 144°, and 162°) and 3 (0°, 54°, and 144°), respectively.

Figure 4.18(a) shows the accuracy from using 2D whole body feature that extracted by MediaPipe. After reduce $|\mathbb{D}|$ to 5 and 3, accuracy from approach 1 is reduced by 15.20% and 37.5%, respectively. Meanwhile, the accuracy of approach 3 reduced by 17.57% and 29.72%. This indicates that the accuracy from each D of approach 3 is higher than approach 1, makes it become more robust than approach 1 when reduces $|\mathbb{D}|$.

In contrast to the accuracy from implementing 3D feature, as in Figure 4.18(b). In this case, the accuracy of approach 1 is reduced by 15.15% and 28.79%, while approach 3 is reduced by 29.9% and 43.28%. It suggests that approach 1 become more robust than approach 3 in this situation. For these results, we are assured that approach 1 requires 3D features. It shows that the 3D feature is more robust than the 2D feature. It suggests that even though the estimated z-axis from MediaPipe might not be as accurate as the stereo vision technique, it can improve the robustness and reliability of the extracted joints, resulting in more consistent results from using these features. However, 2D feature is suitable to apply with approach 3.

4.3.2 Surveillance scenario

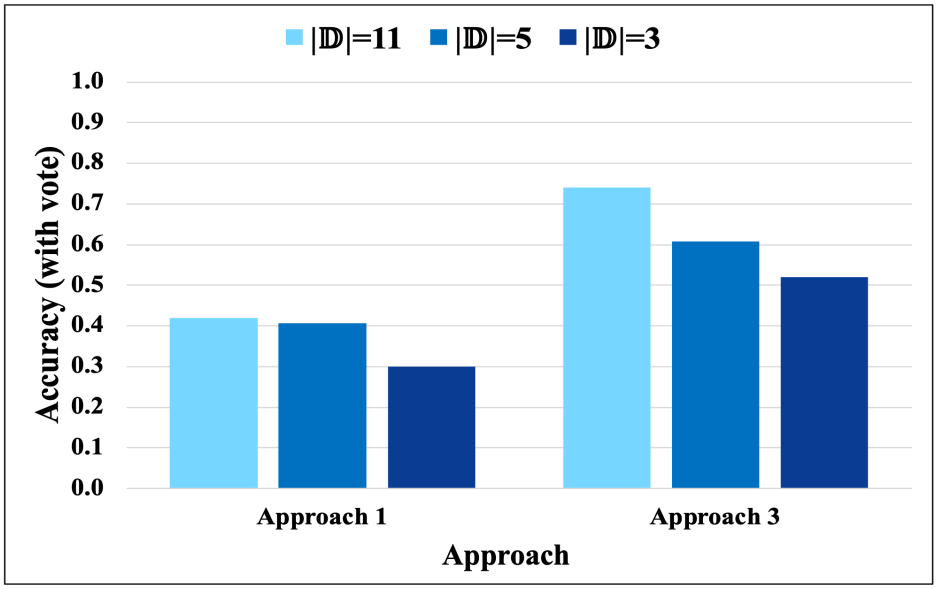
For this scenario, we reduce the number of $\mathbb{D}$ from 14 to 7, where $\mathbb{D}$ = 0°, 30°, 60°, 90°, 195°, 225°, and 255° are selected.

Figure 4.19(a) presents the accuracy from OUMVLP-Pose dataset that

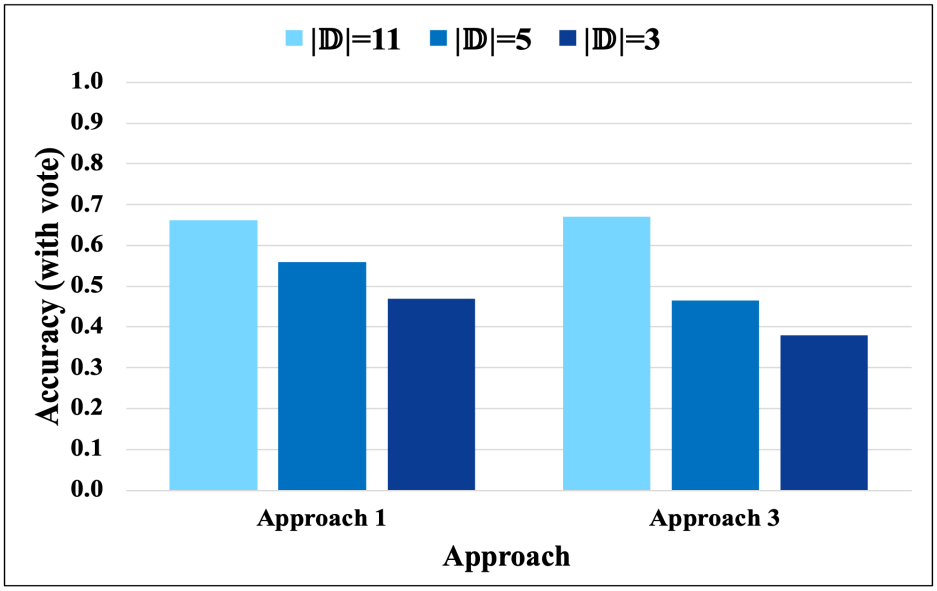
applied OpenPose as a joint estimator. After reduced $|\mathbb{D}|$ to 7 and 5, accuracy of approach 1 reduced by 25% and 33.33%, respectively. While the accuracy of approach 3 is reduced by 21.28% and 48.94%. Even though the accuracy from both approaches become exact value, but the reduction rate of approach 1 is less than approach 3. It suggests that approach 1 is more robust to this variation in this case.

Figure 4.19(b) shows the accuracy of OUMVLP-Pose dataset that used AlphaPose as a pose estimator. We found that the reduction rate of approach 1 after reduced $|\mathbb{D}|$ to 7 and 5 are 20.76% and 32.08%, while approach 3 are 20% and 35%. The result implies that approach 1 is more robust to this variation according to lower reduction rate, even though value is slightly less than approach 3.

Unfortunately, we cannot discuss the result from 3D feature in this scenario due to the original data provided only 2D joints in x-axis and y-axis without z-axis. However, the estimators itself are able to extract 3D joints.

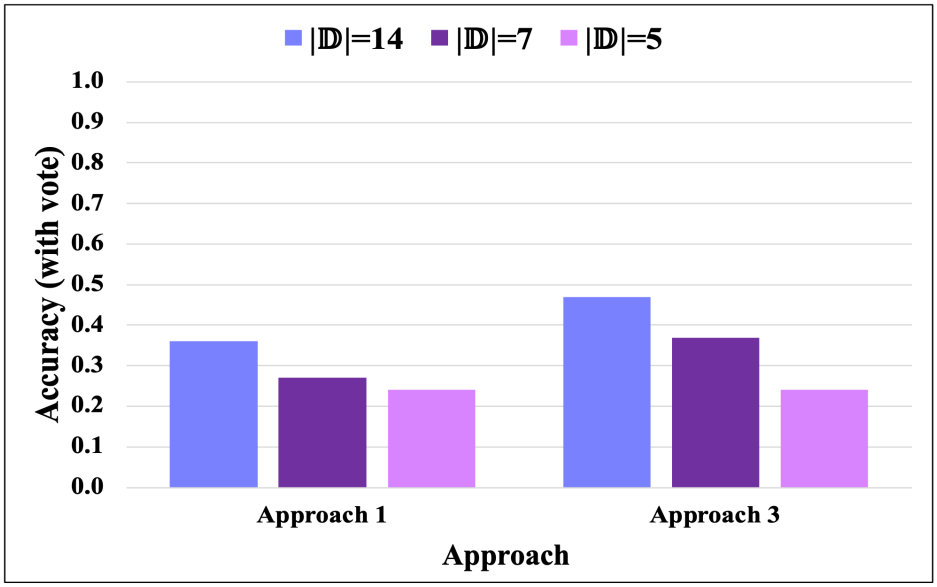


(a)

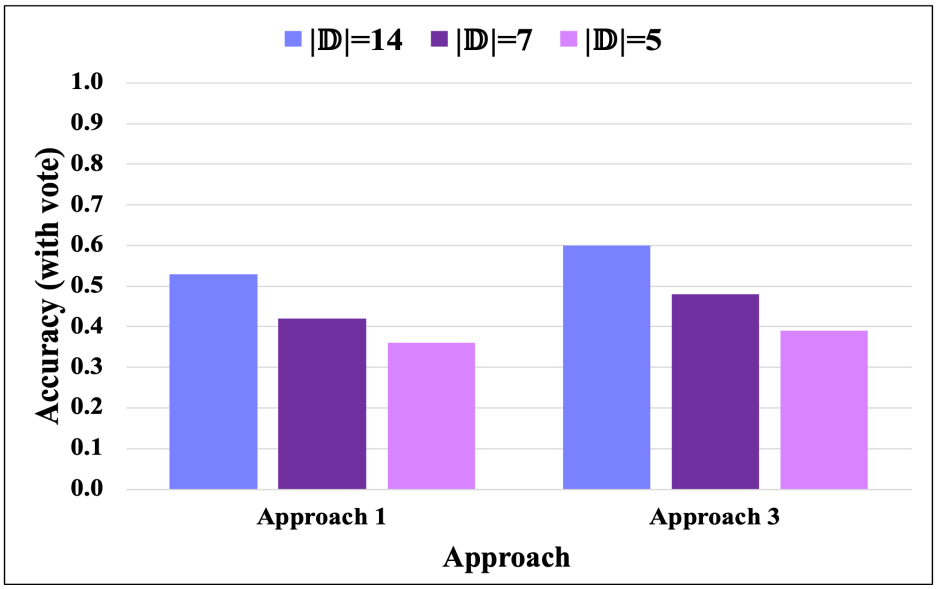


(b)

Figure 4.18: Accuracy of the matching with majority vote on NM sub-dataset from employing whole body joints (2D) when $|\mathbb{D}| = 11$, 5, and 3. (a) 2D joints extracted by MediaPipe. (b) 3D joints extracted by MediaPipe.



(a)



(b)

Figure 4.19: Accuracy of the matching with majority vote on OUMVLP-Pose dataset when $|\mathbb{D}| = 14$, 7, and 5. (a) Joints extracted by OpenPose. (b) Joints extracted by AlphaPose.

4.4 The significance of different joints determination

In this experiment, we will remove each pair of joints one-by-one from the feature vectors. For example, we would like to remove a pair of elbow angles from Equation (3.9) to observe its significance. The feature vector after removing a pair of elbow angles will be $\theta^{i,t,D} = \left[\theta^{3,i,t,D}...\theta^{j,i,t,D}...\theta^{u,i,t,D}\right]^T$. Additionally, we will use the feature vector after removing the pair of desired joints to further calculate the correlation feature.

Notably, we calculate the difference between baseline percentages and the percentage after removing joints by using Equation 4.1.

$$\text{Difference} = \frac{Acc_{removed} - Acc_{baseline}}{Acc_{baseline}} \times 100 \tag{4.1}$$

Where $Acc_{baseline}$ is the accuracy from whole body features of 118 subjects, and we let it be a baseline accuracy. The $Acc_{removed}$ represents the accuracy after removing the desired joints of 118 subjects.

4.4.1 Approach 1: Apply Dynamic Time Warping (DTW) with time-dependent features.

- *Eye-level scenario*

In this scenario, the results from 2D joints in table 4.1 and 3D joints in table 4.2 indicate that the upper body is significant for identifying the identity. It is because most of the features used in this study belong to the lower part, and even if one of them is removed, they associate with each other. However, the upper body also plays a crucial role in this particular scenario. Hence, a removed elbow will lose significant information needed to distinguish a person.

Furthermore, the back ankle appears to be the least significant component. The results show that removing the back ankle can increase the accuracy by 12% and 2.6% from the base line by using 2D and 3D joint coordinates, respectively. This is because when performing a matching, the back ankle is the least changing angle that may be treated as noise, so removing this part leads to an increase in accuracy. Unfortunately, we are unable to identify the exact reason due to insufficient data for experimenting because we have only the CASIA-B dataset.

For this scenario, we arrange the ranking order of the most significant to least significant joints as follows:

1. Elbow
2. Front ankle
3. Knee
4. Hip
5. Back ankle

Table 4.1: Accuracy with majority vote on NM sub-dataset of CASIA-B after removing each pair of joint angle (MediaPipe 2D).

Baseline	Elbow	Hip	Knee	Ankle (f)	Ankle (b)
48%	40%	47%	45%	42%	60%
Difference	-16.67%	-2.08%	-6.25%	-12.5%	+25%

Table 4.2: Accuracy with majority vote on NM sub-dataset of CASIA-B after removing each pair of joint angle (MediaPipe 3D).

Baseline	Elbow	Hip	Knee	Ankle (f)	Ankle (b)
66%	59%	64%	65%	62%	69%
Difference	-10.61 %	-3.03 %	-1.52 %	-6.06 %	+4.55 %

- *Surveillance scenario*

According to the results shown in tables 4.3 and 4.4, the hip part is the most significant in this scenario. It implies that the hip is the most accurate feature to make patterns more distinguishable. It might be because of the hip located at the center of the body, and even though the perspective of the camera changed, it does not affect much. In fact, knees and elbows are also important, but in this scenario, it easily caused the wrong estimation of their location. The result suggests that the error from estimating the elbow of OpenPose is higher than that of Alphapose, which agrees with the result shown in Figure 4.9 that the upper part feature extracted by OpenPose is unreliable.

In this scenario, we rank the most significant to least significant joints based on the OpenPose estimator as follows:

1. Hip
2. Knee
3. Elbow

Additionally, the ranking order of the joints based on AlphaPose are ordered as follow:

1. Hip
2. Elbow
3. Knee

Table 4.3: Accuracy with majority vote on OUMVLP-Pose dataset after removing each joint angle (OpenPose).

Baseline	Elbow	Hip	Knee
36%	47%	32%	37%
Difference	+30.56 %	-11.11 %	+2.78 %

Table 4.4: Accuracy with majority vote on OUMVLP-Pose dataset after removing each joint angle (AlphaPose).

Baseline	Elbow	Hip	Knee
53%	48%	47%	53%
Difference	-9.43 %	-11.32 %	0 %

4.4.2 Approach 3: Apply Euclidean distance (EU) with time-independent feature.

- *Eye-level scenario*

In this scenario, tables 4.5 and 4.6 show that the most significant part is the hip, and the least significant part is the back ankle on both 2D and 3D joints from the MediaPipe estimator. It indicates that the hip is a center part that connects each joint together, as the correlation will connect joints by

finding the relationship between them. If we remove the hip, the relationship between the remaining parts becomes insignificant, making it impossible to identify individuals.

Moreover, the upper body part is an important part in this scenario, but it is not the first priority for the time-independent correlation feature. It is because this feature treats the upper body as a support for the lower body. The other lower parts achieve lower significance than the elbow because most of them belong to the lower body, which makes them associate with each other. Additionally, the results suggest that it is a better choice to not include the back ankle. Eliminating it can enhance the precision of both approaches.

In this scenario, the ranking order for the most significant to least significant joints is as follows:

1. Hip
2. Elbow
3. Front ankle
4. Knee
5. Back ankle

Table 4.5: Accuracy with majority vote on NM sub-dataset of CASIA-B after removing each pair of joint angle for calculating the time-independent correlation (MediaPipe 2D).

Baseline	Elbow	Hip	Knee	Ankle (f)	Ankle (b)
74%	58%	48%	64%	61%	79%
Difference	-20.99%	-34.73%	-14.11%	-17.54%	+6.49%

- *Surveillance scenario*

The result from OpenPose in table 4.7 suggests that hip is the most important part for this estimator, as the accuracy is drastically decreased after removing it. Moreover, this result indicates that the elbow, which is a representative of the upper body, is supporting the lower body. Removing the elbow resulted in a 30% decrease in accuracy. This result related to Figure **??** shows that only the lower body is insufficient for identifying people. However, the knee

Table 4.6: Accuracy with majority vote on NM sub-dataset of CASIA-B after removing each pair of joint angle for calculating the time-independent correlation (MediaPipe 3D).

Baseline	Elbow	Hip	Knee	Ankle (f)	Ankle (b)
66%	59%	64%	65%	62%	69%
Difference	-10.61 %	-3.03 %	-1.52 %	-6.06 %	+4.55 %

appears to be the least significant part when using OpenPose. It is because the lower body consists of parts, and the hip takes more weight than the knee. Although the knee is the least significant part, its inclusion is crucial as its removal results in a decrease in accuracy.

The result from using AlphaPose in table 4.8 indicates a difference. It implies that the knee is the most important part, followed by the hip and the elbow. This means the knee is the main part that connects each joint together. Furthermore, the elbow serves as a vital support for the lower body. It achieves the least significance but does not suggest being discarded. It is because the accuracy is decreased by 27% after the elbow is removed.

In this scenario, the ranking order for the most significant to least significant joints based on OpenPose is as follows:

1. Hip
2. Elbow
3. Knee

Additionally, the ranking order for the most significant to least significant joints based on AlphaPose is as follows:

1. Knee
2. Hip
3. Elbow

Table 4.7: Accuracy with majority vote on OUMVLP-Pose dataset after removing each pair of joint angle for time-independent correlation calculation (AlphaPose).

Baseline	Elbow	Hip	Knee
47%	17%	15%	21%
Difference	-63.38 %	-68.09 %	-55.32 %

Table 4.8: Accuracy with majority vote on OUMVLP-Pose dataset after removing each pair of joint angle for time-independent correlation calculation (AlphaPose).

Baseline	Elbow	Hip	Knee
60%	33%	30%	28%
Difference	-45 %	-50 %	-53.33 %

4.5 Comparative results of different voting algorithms

This section will discuss on the comparison between majority vote and weighted vote to find a proper voting algorithm for our method. Notably, the time-dependent correlation feature fails to identify people due to insufficient data variations. Neither a majority vote nor a weighted vote can improve the results from this feature. Hence, we will discuss only the results from approach 1 and approach 3.

Figure 4.20 presents a methodology of our weighted vote. After matching, the weighted voting is simply starting from employ the accuracy of each D (Acc^D) to be its weight as w^D by the following equation:

$$w^D = \frac{Acc^D}{\sum Acc^D} \tag{4.2}$$

Next, calculate the score of each D, called $score^{i_k^D}$, by a following equation:

$$score^{i_k^D} = w^D \times p \tag{4.3}$$

where p is a binary score that refer to true or false as shown below:

$$p = \begin{cases} 1, \text{if correct } i_k^D \\ 0, \text{otherwise} \end{cases} \tag{4.4}$$

After that, we find the total score of every perspectives as:

$$score^{i_k} = \sum score^{i_k^d} \tag{4.5}$$

Finally, the matched person (i_k) will be selected by maximum $score^{i_k}$ as follows:

$$i_k = \max(score^{i_k}) \tag{4.6}$$

The results in Figures 4.21–4.22 present the comparative results between majority vote and weighted vote on both approaches. Notably, the threshold for weighted vote is set to 0.1, which produces the best results. It indicates that the weighted voting algorithm is unsuitable to apply with our method in both scenarios. It is because it assigns weights that depend on the accuracy of each camera. If the Acc^D is high enough, this voting technique is efficient. However, the Acc^D produced by our proposed method varies depending on each D. Some perspectives achieve very high Acc^D but fail on the others. By this reason, threshold required to be small, but it inefficient.

For this reason, we select a simple majority vote to integrate as much information as possible to enhance the accuracy of the matching. In fact, the majority vote is not the best algorithm, but at the initial point, it proved to be more suitable than complicated weighted voting.

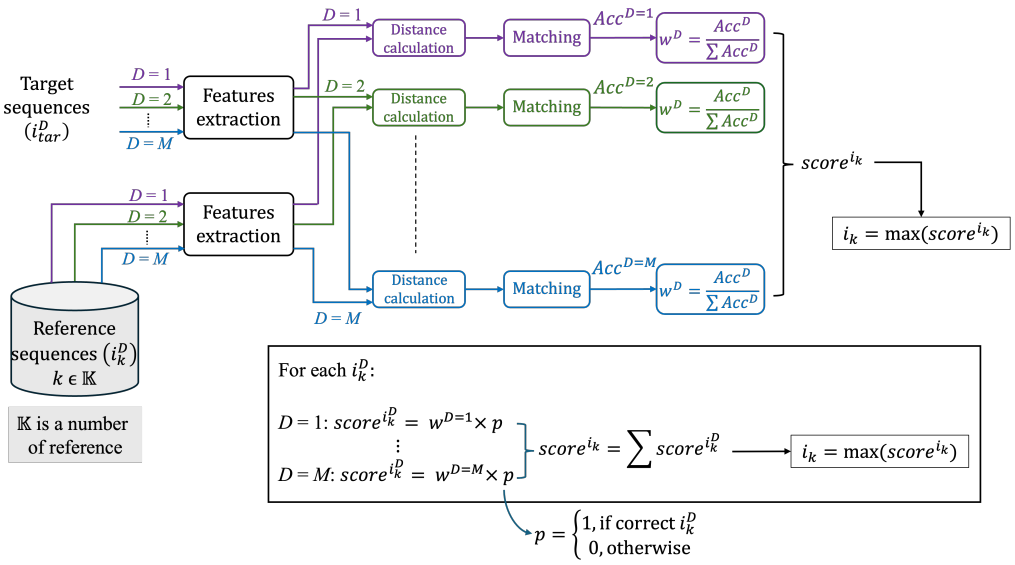
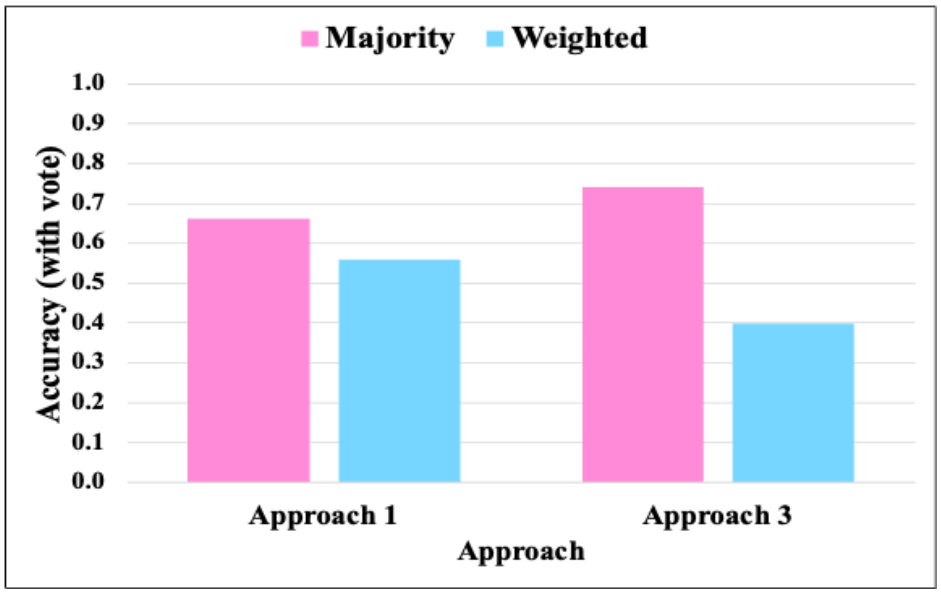
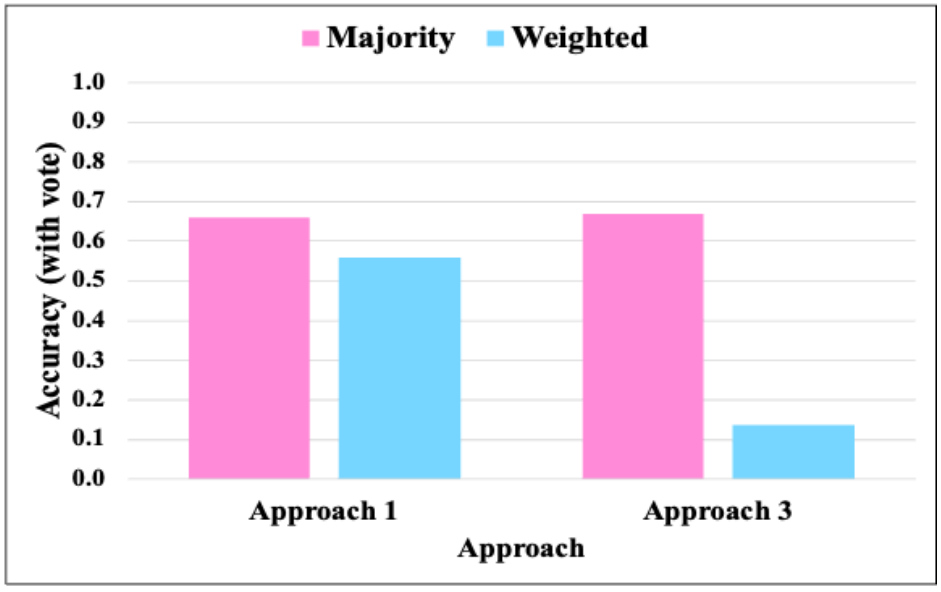


Figure 4.20: Diagram to describe the methodology of weighted voting.

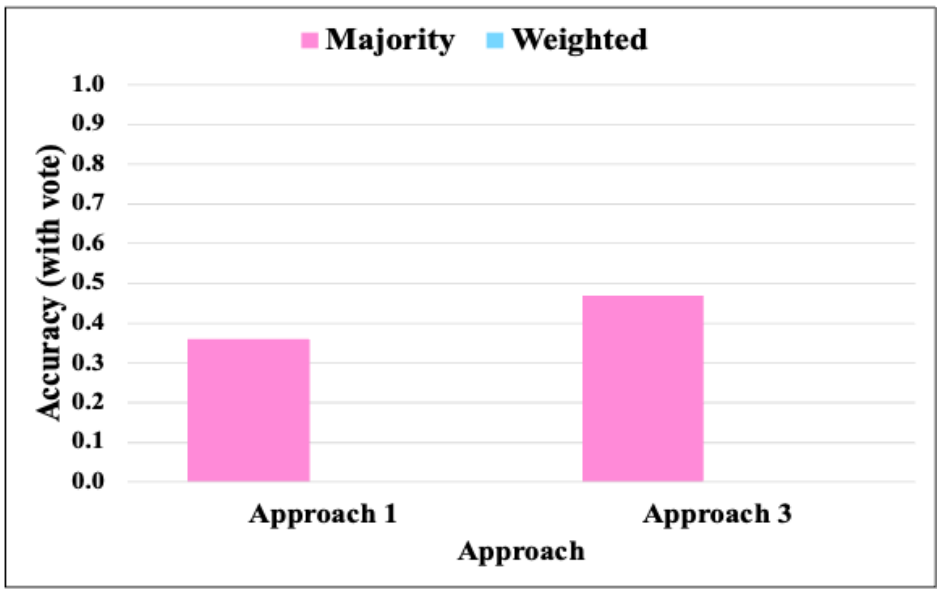


(a)

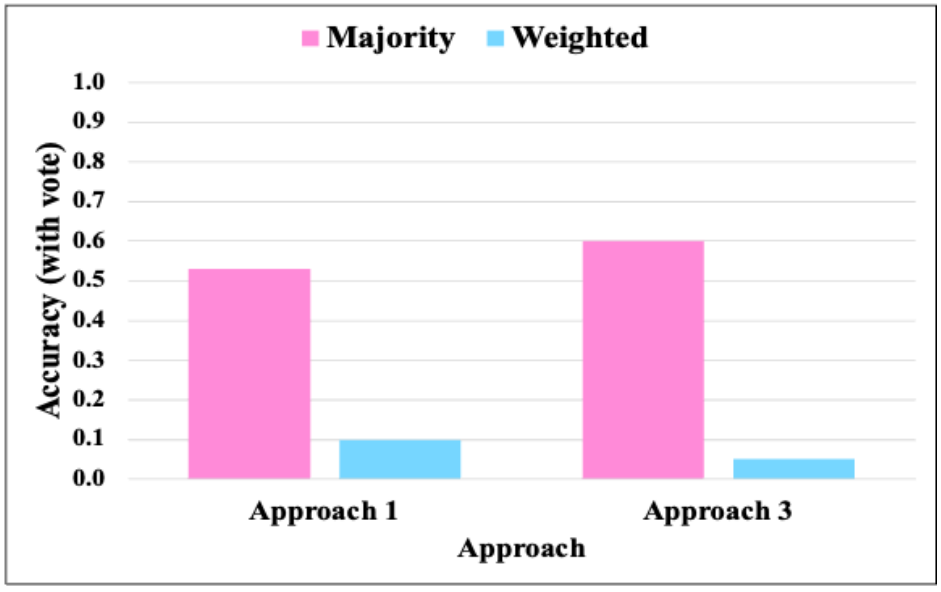


(b)

Figure 4.21: Accuracy of the matching with majority vote on NM sub-dataset. These results are from employed 2D joints extracted by MediaPipe. (a) Accuracy with majority vote from employing 2D joints. (b) Accuracy with majority vote from employing 3D joints.



(a)



(b)

Figure 4.22: Accuracy of the matching with majority vote on OUMVLP-Pose dataset. (a) Accuracy with majority vote from employing 2D joints by OpenPose. (b) Accuracy with majority vote from employing 2D joints by AlphaPose.

4.6 Comparative results between distance measurement algorithms

In this experiment, we compare the DTW and EU methods to determine which one is most suitable for walking pattern matching. As described in Chapter 3, approach 1 employs DTW to match the time-dependent features, which are joint angles and time-dependent correlation. However, insufficient data variations prevent the application of time-dependent correlation for pattern matching. Hence, we will employ only joint angles. For approach 3, the EU employs the time-independent correlation feature. We will use the exact video sequence for matching, one will be an original sequence, and the other will be the same sequences with the starting time shifted by 0, 5, 10, 15, and 20 frames.

4.6.1 Eye-level scenario

The experiment conditions of this scenario are as following:

- The average steps per sec. = 1.47 steps/sec.
- The frame rate = 8 frames/sec.
- Number of step per frame = 0.183 step/frame

Hence, the step shifted conditions will be:

- 5 frames shifted = 0.92 step shifted.
- 10 frames shifted = 1.83 step shifted.
- 15 frames shifted = 2.75 step shifted.
- 20 frames shifted = 3.66 steps shifted.

Figures 4.23a and 4.23b show the results with majority vote of the matching between original sequences and the same sequences with time shifted. This allows for a comparison between approach 1 (DTW with joint angles feature) and approach 3 (EU with time-independent correlation feature).

This scenario enhances gait information by clearly displaying the entire body. When compared to exact sequences, the time delay has a minimal impact. Still, results show that the performance of approach 3 is significantly reduced when we increase the frame shift. Figure **??**b shows that the accuracy of approach 3 is starting to reduce when time is shifted by 15 frames.

It implies that approach 1 is better at handling this problem, which suggests that approach 1 performs better when a person walks at a different speed and starting position. Meanwhile, approach 3 may treat the exact person as others when their walking speed and starting position are changed, resulting in a misidentification. However, both approaches are not much different when we apply a majority vote, as it will enhance the performance of the matching results, but the accuracy of approach 1 is still higher than approach 3 when shifting 20 frames. Moreover, approach 1 performs better without specifying the starting point and normalizing the time of the data, unlike approach 3, which requires time normalization by rearranging the data from the entire sequence and reordering it according to its ranks instead of using the original feature that includes time information.

4.6.2 Surveillance scenario

The experiment conditions of this scenario are as following:

- The average steps per sec. = 1.47 steps/sec.
- The frame rate = 25 frames/sec.
- Number of step per frame = 0.059 step/frame

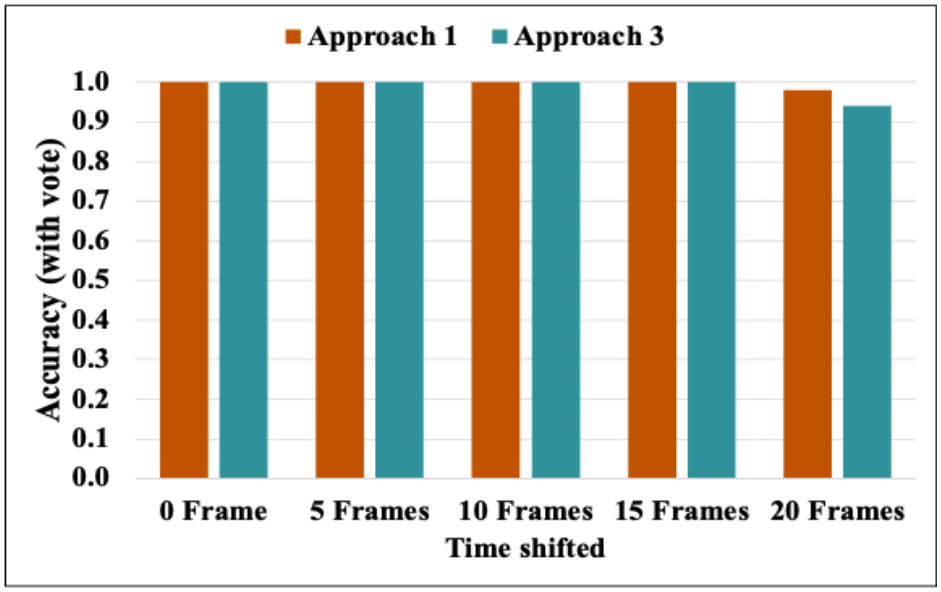
The step shifted conditions for this scenario:

- 5 frames shifted = 0.29 step shifted.
- 10 frames shifted = 0.59 step shifted.
- 15 frames shifted = 0.88 step shifted.
- 20 frames shifted = 1.18 steps shifted.

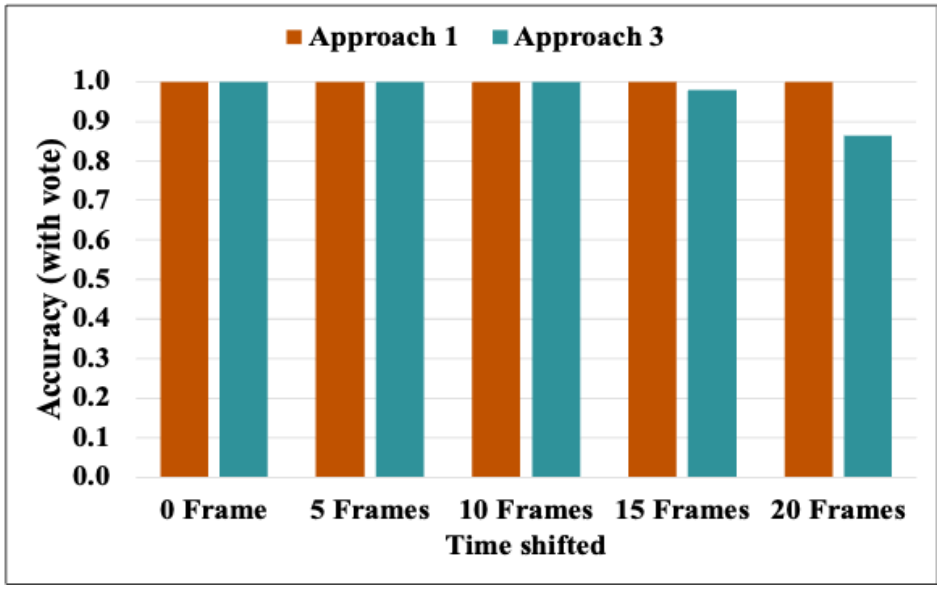
Figure 4.24a shows the comparative results of delaying the starting time for 0, 5, 10, 15, and 20 frames between approaches 1 and 3 that employ the features based on the estimated joints from OpenPose. Furthermore, Figure 4.24b shows the comparative results from employing the features based on the AlphaPose estimator.

The results from this scenario confirm that approach 1 is the best at handling the time series data, especially for 15 frames. Despite the uneven pattern lengths and significant reduction in matching information, approach 1 outperforms approach 3. The 20 frames shifted imply we remove almost the entire sequence, the accuracy is dramatically reduced on both approaches. It

suggests that approach 3 performs best when everyone is walking at the same speed. However, this scenario is ideal, indicating that approach 3 is more suitable for application under controllable conditions, unlike in a surveillance setting. These results show that approach 1 is the best at serving the walking pattern in a real-world scenario where data normalization is a challenge and it is impossible to specify the starting point, which can enhance the system's flexibility.

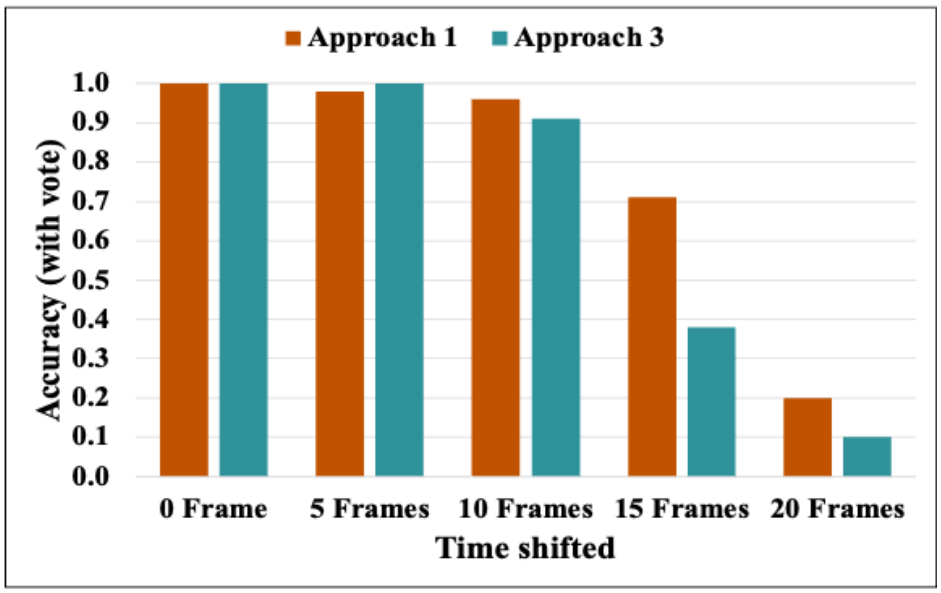


(a)

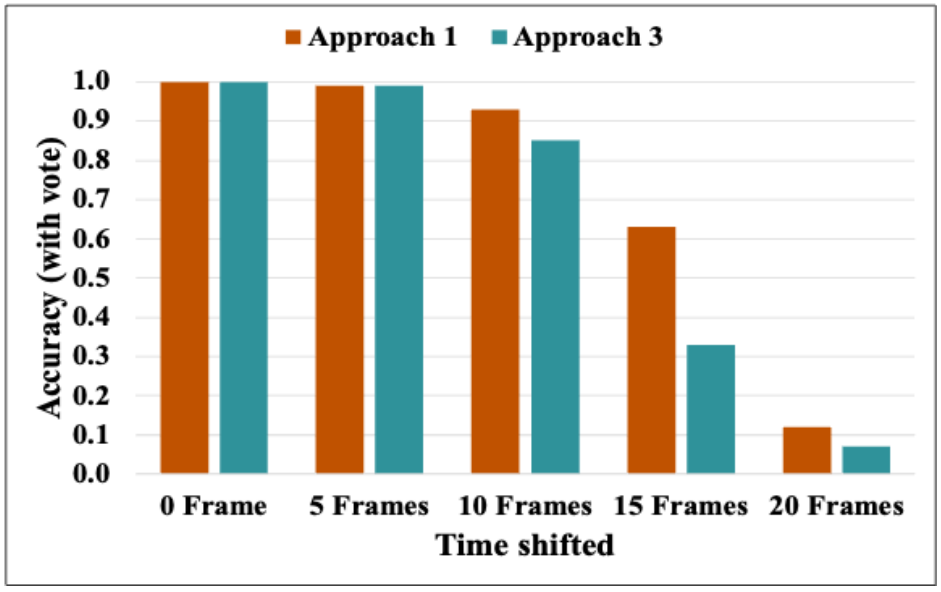


(b)

Figure 4.23: Accuracy of the matching after shifting the time with majority vote on NM sub-dataset. (a) Result from employing 2D joints. (b) Result from employing 3D joints.



(a)



(b)

Figure 4.24: Accuracy of the matching after shifting the time with majority vote on OUMVLP-Pose dataset. (a) Result from employing 2D joints extracted by OpenPose. (b) Result from employing 2D joints extracted by AlphaPose.

4.7 Comparative results with prior studies

This section shows comparative results with existing studies. We select each of the appearance-based and model-based approaches for comparison. The appearance-based approach requires more complicated features than the model-based approach, and it achieves an impressive result. Meanwhile, the model-based approach requires a more complex model but employs simpler features, such as joints from skeleton landmarks. It achieves a competitive result but is still lower than the appearance-based approach. We then select both approaches for comparison to study the process and the differences between them and ours. The CSTL [24] is a representative of the appearance-based approach, and GaitGraph2 [50] is a representative of the model-based approach, which is employed to analyze the gait with the same condition as ours.

We let NM01 be the training set and NM02 be the test set. The gallery and probe in the test set will have the same data as ours, as we do not separate the test set into gallery and probe. Our method will match the data in NM02 with the data in NM01 to select the matched person. We either separate the number of subjects into 20, 49, and 118 subjects for both the training and test sets. For the 20 subjects, we train 20 subjects of NM01 and test with 20 subjects of NM02, and do the same with 49 and 118 subjects. For the other parameters, we use the original values as described in their papers and specified in the code, but we adjust the iteration to be 20,000 and 2,000 for CSTL and GaitGraph2, respectively. The original iteration for CSTL was 100,000, and the GaitGraph2 iteration was 500. We selected the result from approach 1 using the whole body joint angle feature (without the back ankle) based on the 3D joints from MediaPipe for comparison, the approach 2 uses the results based on 2D joints from MediaPipe. All results from both approaches are the best case for each.

Table **??** shows the comparative results between CSTL, GaitGraph2, and ours. It shows that our approach 1 performs better than CSTL on 49 and 118 subjects, and our approach 2 method achieves higher accuracy than CSTL. It suggests that our method has higher recognition performance on the smaller dataset. We need two sub-datasets to serve as a reference and target for matching, which are similar to the training and testing sets for CSTL. Furthermore, accuracy dropped significantly when the number of subjects changed. It indicates that CSTL requires more data and higher computational time to tune their model for consistency and impressive results, as presented in their paper. Additionally, the appearance-based approach lacks structural features and contains irrelevant gait features that may mislead the model, making it unsuitable to apply in a real-world situation.

When compared to our approach 1, GaitGraph2 achieves the highest accuracy. However, the accuracy of 49 subjects from Approach 2 is slightly higher than GaitGraph2. They utilized the multi-branches features, i.e., joints, motion, and bones, to recognize people, but we use the joint angles feature for identification. This suggests a variation in the number of features, and it excludes the process of feature learning. However, there is a slight gap between ours and GaitGraphs2. Furthermore, our method provides more flexibility to the users, especially when they wish to make changes to the quantity in the database, such as deleting or adding people.

Additionally, tables 4.10 and 4.11 present the comparative results of GaitGraph2 and ours on the OUMVLP-Pose dataset. This dataset provides the joint coordinates that were extracted by OpenPose and AlphaPose. The results show that our method outperforms GaitGraph2 on this dataset, especially for Approach 2. Approach 1 achieves competitive results when employing the joints extracted by OpenPose, and it outperforms GaitGraph2 on 20 and 100 subjects when utilizing the joints extracted by AlphaPose. In fact, this dataset only allows us to calculate six angles, i.e., the elbow, hip, and knee on the left and right sides, because it lacks a foot landmark. Thus, we lost the one piece of information that is crucial for walking patterns, making identification from three parts of the body more challenging. However, the result shows our method has great potential for application in surveillance scenes.

Approach 2 of our method appears to have higher accuracy than Approach 1, but its consistency is lower. Consequently, increasing the number of subjects significantly reduces the accuracy, in contrast to approach 1, which effectively handles this variation. It implies that approach 2 is more sensitive to outliers and noises than approach 1. Moreover, approach 1 can better handle the walking speed variation than approach 2. By the way, these results prove that pattern matching has an impressive ability to perform gait analysis for identification, especially when a small amount of data is available. In addition, it necessitates a non-complex environmental condition to perform this task, unlike DNNs. It suggests that pattern matching is an affordable alternative method for accessing the identification task.

Table 4.9: Comparative rank-1 accuracy on NM sub-dataset of the CASIA-B from CSTL [24], GaitGraph2 [50], and ours.

	Number of subjects		
	20	49	118
CSTL	86.09%	72.38%	65.09%
GaitGraph2	91.17%	87.18%	82.00%
Ours (Approach 1)	80.00%	74.00%	69.00%
Ours (Approach 2)	90.00%	88.00%	74.00%

Table 4.10: Comparative rank-1 accuracy on OUMVLP-Pose (OpenPose) from GaitGraph2 [50] and ours.

	Number of subjects		
	20	50	100
GaitGraph2	59.04%	52.21%	46.87%
Ours (Approach 1)	55.00%	52.00%	47.00%
Ours (Approach 2)	80.00%	64.00%	47.00%

Table 4.11: Comparative rank-1 accuracy on OUMVLP-Pose (AlphaPose) from GaitGraph2 [50] and ours.

	Number of subjects		
	20	50	100
GaitGraph2	62.68%	62.15%	52.61%
Ours (Approach 1)	65.00%	62.00%	53.00%
Ours (Approach 2)	75.00%	70.00%	60.00%

Chapter 5

Conclusion & Future works

5.1 Contributions

Gait is an individual walking pattern that is established by the changing of body joints over a period of time. When each joint changes its position, the posture is noticeable. We all have different postures while walking, depending on individual walking speed, arm swing, foot placement, weight transfer, and so on. It is related to the neurological control that expresses our walking trait. Basically, gait presents transportation information, e.g., walking direction and predicted destination. On top of that, gait represents insight into individual information, such as age, gender, activity, emotion, health condition, personality, and identity.

Vision-based gait analysis is a system that analyzes the gait based on images or videos. It requires no contact with walkers, making it a distance analysis system. Both single-view and multi-view gait analyses extensively use the camera as a tool to access gait information due to its simplicity, scalability, flexibility, and cost-effectiveness. However, it encounters various challenges, such as sensitivity to environmental factors, dynamic backgrounds, occlusions, and view variation, especially in a multiple surveillance camera environment. For the identification task, view variation affects misidentification that is caused by changes in camera perspectives.

This study presents multi-view gait recognition that aims to integrate the data from multiple cameras by a majority vote based on the features of human body parts. Our purpose is different from the existing studies as we aim to identify the known person from multiple perspectives, and DNNs are unnecessary for this task. Hence, only the pattern matching that is supervised by the reference data in the database is sufficient.

We test the experiment on the CASIA-B and OUMVLP-Pose datasets.

The CASIA-B dataset is representative of the eye-level scenario, which is a situation when cameras are equipped at the same level as human eyes. Meanwhile, the OUMVLP-Pose dataset provides scenes from a higher position, similar to a surveillance scene.

We analyze the gait by determining joint angles and their correlation, which represents the motion of walking in the sequences. There are two correlation features, i.e., time-dependent correlation and time-independent correlation. Additionally, we propose two approaches according to the features to study the human gait with and without time information. Approach 1 utilizes DTW with joint angles and time-dependent correlation features, and Approach 2 applies EU with the time-independent correlation feature.

We divide features into three parts, i.e., whole, upper, and lower body, to study the impact of different body parts on gait analysis. Additionally, we removed each joint one by one to study its importance to the gait analysis. Then, employ these features to match people in separate multiple cameras. Finally, apply a majority vote to integrate the separated data to improve accuracy. Furthermore, we divide the number of subjects into 20, 50, and 118 subjects for the CASIA-B dataset and 20, 50, and 100 subjects for the OUMVLP-Pose dataset to observe the trend of the matching accuracy when the number of subjects is varied.

According to the findings, integrating the view variations by majority voting can improve the view-variation of multi-view gait analysis. We found that the upper body part of the joint angles feature is essential for the eye-level scenario in addition to the lower body. Without upper body feature, leads to decrement of accuracy. Notably, the back ankle angles that includes in a feature vector of eye-level scenario should be treated as a noise. The results indicate that remove it can increase the accuracy of either approach 1 and approach 3. However, the joint angles feature related to the lower body are sufficient for identifying people in the surveillance scenario when OpenPose is applied. For the case of AlphaPose, whole body feature is the best. Unfortunately, the approach 2 that used time-dependent correlation fails to identify people due to insufficient data variations.

We found that approach 3, which applies EU with the time-independent correlation feature that calculates the correlation between joint angles of the entire sequences, requires features from the whole body part. Basically, correlation needs variations of data in order to specify the relationship between them. As a result, the whole body feature, which includes both the upper and lower body, is the best feature for representing the walking pattern using the time-independent correlation. The weakness of approach 3 is that it is unsuitable to apply with time series data, such as a walking pattern of the same person with different speeds. This causes a change in the calculated

correlation value and leads to a misidentification. It makes approach 3 unsuitable for applying to an uncontrollable environment, such as a surveillance scenario, in contrast to approach 1. Moreover, approach 1 is flexible with a non-normalized data, which makes it more practical to apply with real-world scenarios that challenges on data normalization issue.

The experimental results suggest that the proposed method is suitable for identifying identities with a small quantity of databases, not only the people quantity but the sequences. Since we employ two sub-datasets per each of the CASIA-B and OUMVLP-Pose datasets, We achieve a competitive result when compared to state-of-the-art methods. Our method is more flexible when changing the number of databases due to a non-training method, e.g., when adding or deleting people. Additionally, the availability of data visualization enables one-by-one detailed analyses, which will be advantageous for the expansion of our future tasks. Furthermore, it can be executed on the CPU according to its non-training state. Thus, the GPU and complicated environment are unnecessary, leading to reductions in both cost and time.

5.2 Addressing the research questions

- *RQ1: How to analyze human motion from a multi-view gait image for human behavior analysis based on their walking pattern?*

We propose a pattern matching method based on temporal geometric features of human body parts.

- *RQ2: How to improve human gait analysis method from the multi-view gait image sequences for person identification under surveillance scenarios?*

We improve the view-variation issue of human gait analysis method by applying majority vote to integrate the analysis from multiple perspectives together.

- *RQ3: How to explore the optimal feature to estimate the human gait?*

We conduct a comparative study between the different joint features.

5.3 Limitations & Future works

5.3.1 Short-range plans

- **Improve the matching algorithm**

The measured DTW distance shows that the shortest distance between patterns can be different, resulting in a mismatch and reducing accuracy. In the case of mismatch, we observe that there is mostly a slight difference in the distance between true-matched (an identity that should be selected) and false-matched (the person that mismatches with the true label) identities. Thus, modifying the matching algorithm can improve matching accuracy. Furthermore, it can improve the matching algorithm by handling a larger number of subjects, which is a significant issue that reduces accuracy.

- **Improve the occlusions problem**

Mostly, the model-based gait analysis relies on the pose estimation method. In fact, the existing algorithms can handle the occlusion problem, but they may be inaccurate. This makes the gait analysis system produce inaccurate results, especially when most of the body is occluded. This problem is another challenge for the vision-based gait analysis that requires improvement.

- **Estimate the gait cycle based on a walking pattern**

Since each walking pattern contains many gait cycles, this cycle holds essential individual information. It can be an effective feature for identifying the identity and handling the unequal walking speed and length of the same person's sequences, which affects the mismatch in time-series data. Even if the DTW distance can handle these issues, the starting point needs to be identical. It indicates that the captured sequences require the same starting point with an identical movement pattern of joints for the most efficient matching based on DTW. However, this situation is impossible in real-world situations because we cannot control the starting point and movement of walkers. Hence, the gait cycle from the walking pattern is a key to improving matching based on DTW.

- **Extension the gait analysis tasks**

As discussed in Chapter 1, gait serves a variety of purposes, including activity, clinical, and emotion analysis. To achieve these objectives, we need to analyze the gait using a classification method. Then, DNN plays a significant role in achieving it because only pattern matching is insufficient to analyze and obtain such information.

5.3.2 Long-term visions

- **View-free gait analysis**

In a real-world situation, we cannot specify the direction in which people walk because they can walk randomly from anywhere, unlike in a laboratory environment, where we can control every variable and parameter as required. In fact, multi-view may not be sufficiently practical to address every situation in a real-world setting. In this case, view-free gait analysis is key to handling real-world situations. We expect the view-free gait analysis to function as a 3D model of human gait, applicable to all perspectives. To achieve this, it consumes money and time, and it needs more data to study the gait.

- **All-in-one gait analysis system**

After completely addressing the gait analysis for every task, we can aggregate them into one system, whose result depends on the specified purpose. It will have a great positive impact and benefit us by serving a complete system that can improve clinical, psychological, security, and more.

Bibliography

[1] Munif Alotaibi and Ausif Mahmood. Improved gait recognition based on specialized deep convolutional neural network. *Computer Vision and Image Understanding*, 164:103–110, 2017.

[2] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):421–430, 2020.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[4] Danilo Avola, Luigi Cinque, Maria De Marsico, Alessio Fagioli, Gian Luca Foresti, Maurizio Mancini, and Alessio Mecca. Signal enhancement and efficient dtw-based comparison for wearable gait recognition. *Computers & Security*, 137:103643, 2024.

[5] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

[6] Yajurv Bhatia, ASM Hossain Bari, and Marina Gavrilova. A lstm-based approach for gait emotion recognition. In *2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, pages 214–221, 2021.

[7] Michalina Błażkiewicz, Karol Lann Vel Lace, and Anna Hadamus. Gait symmetry analysis based on dynamic time warping. *Symmetry*, 13(5), 2021.

[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[10] Kuan-Yu Chen, Jungpil Shin, Md. Al Mehedi Hasan, Jiun-Jian Liaw, Okuyama Yuichi, and Yoichi Tomioka. Fitness movement types and completeness detection using a transfer-learning-based deep neural network. *Sensors*, 22(15), 2022.

[11] Anthony Cimorelli, Ankit Patel, Tasos Karakostas, and R James Cotton. Validation of portable in-clinic video-based gait analysis for prosthesis users. *Scientific Reports*, 14(1):3840, 2 2024.

[12] Muqing Deng and Cong Wang. Gait recognition under different clothing conditions via deterministic learning. *IEEE/CAA Journal of Automatica Sinica*, 2018.

[13] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14213–14221, 2020.

[14] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[15] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[16] Y. Fu, S. Meng, S. Hou, X. Hu, and Y. Huang. Gpgait: Generalized pose-based gait recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19538–19547, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society.

[17] Yao Ge, Wenda Li, Muhammad Farooq, Adnan Qayyum, Jingyan Wang, Zikang Chen, Jonathan Cooper, Muhammad Ali Imran, and Qammer H. Abbasi. Logait: Lora sensing system of human gait recognition using dynamic time warping. *IEEE Sensors Journal*, 23(18):21687–21697, 2023.

[18] D.N. Gujarati. *Essentials of Econometrics*. SAGE Publications, 2021.

[19] Jay Gupta, Pushkar Dixit, Nishant Singh, and Vijay Semwal. Analysis of gait pattern to recognize the human activities. *International Journal of Interactive Multimedia and Artificial Intelligence*, 2, 07 2014.

[20] Chidananda H. and Hanumantha Reddy. Human activity recognition using foots movement patterns. *International Journal of Research in Advent Technology*, 7:462–467, 04 2019.

[21] Kun Han and Xinyu Li. Research method of discontinuous-gait image recognition based on human skeleton keypoint extraction. *Sensors*, 23(16), 2023.

[22] Chang Soon Tony Hii, Kok Beng Gan, Nasharuddin Zainal, Norlinah Mohamed Ibrahim, Shahrul Azmin, Siti Hajar Mat Desa, Bart van de Warrenburg, and Huay Woon You. Automated gait analysis based on a marker-free pose estimation model. *Sensors*, 23(14), 2023.

[23] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision*, pages 382–398. Springer, 2020.

[24] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12909–12918, 2021.

[25] Robert Hughes, Fadi Muheidat, Mason Lee, and Lo'ai A. Tawalbeh. Floor based sensors walk identification system using dynamic time warping with cloudlet support. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 440–444, 2019.

[26] Sandra R. Hundza, William R. Hook, Christopher R. Harris, Sunny V. Mahajan, Paul A. Leslie, Carl A. Spani, Leonhard G. Spalteholz, Benjamin J. Birch, Drew T. Commandeur, and Nigel J. Livingston. Accurate and reliable gait cycle detection in parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):127–137, 2014.

[27] Nitchan Jianwattanapaisarn, Kaoru Sumi, Akira Utsumi, Nirattaya Khamsemanan, and Cholwich Nattee. Emotional characteristic analysis of human gait while real-time movie viewing. *Frontiers in Artificial Intelligence*, 5, 2022.

[28] Gu Eon Kang, Brian Mickey, Barry Krembs, Melvin McInnis, and Melissa Gross. The effect of mood phases on balance control in bipolar disorder. *Journal of Biomechanics*, 82, 11 2018.

[29] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[30] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. Multi-view large population gait database with human meshes and its performance evaluation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(2):234–248, 2022.

[31] Xiangying Li, Mingwei Zhang, Junnan Gu, and Zhi Zhang. Fitness action counting based on mediapipe. In *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–7, 2022.

[32] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *Computer Vision – ECCV 2022*, 2022.

[33] Rijun Liao, Zhu Li, Shuvra S Bhattacharyya, and George York. Posemapgait: A model-based gait recognition method with pose estimation maps and graph convolutional networks. *Neurocomputing*, 501:514–528, 2022.

[34] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[36] Alvaro Muro-de-la Herran, Begonya Garcia-Zapirain, and Amaia Mendez-Zorrilla. Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications. *Sensors*, 14(2):3362–3394, 2014.

[37] C. S. Myers and L. R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected-word recognition. *The Bell System Technical Journal*, 60(7):1389–1409, 1981.

[38] Jacquelin Perry and Judith M Burnfield. *GAIT ANALYSIS: NORMAL AND PATHOLOGICAL FUNCTION*. USA: Slack Incorporated, 2010.

[39] Walter Pirker and Regina Katzenschlager. Gait disorders in adults and the elderly: A clinical guide. *Wiener klinische Wochenschrift*, 129, 10 2016.

[40] Ana Patrícia Rocha, Hugo Choupina, Jose Maria Fernandes, Maria Jose Rosas, Rui Vaz, and João Paulo Silva Cunha. Parkinson's disease assessment based on gait analysis using an innovative rgb-d camera system. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3126–3129, 2014.

[41] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[42] Jasvinder Pal Singh, Sanjeev Jain, Sakshi Arora, and Uday Pratap Singh. Vision-based gait recognition: A survey. *Ieee Access*, 6:70497–70527, 2018.

[43] Djordje Slijepcevic, Fabian Horst, Sebastian Lapuschkin, Brian Horsak, Anna-Maria Raberger, Andreas Kranzl, Wojciech Samek, Christian Breiteneder, Wolfgang Immanuel Schöllhorn, and Matthias Zeppelzauer. Explaining machine learning models for clinical gait analysis. *ACM Trans. Comput. Healthcare*, 3(2), dec 2021.

[44] Djordje Slijepcevic, Matthias Zeppelzauer, Fabian Unglaube, Andreas Kranzl, Christian Breiteneder, and Brian Horsak. Explainable machine learning in human gait analysis: A study on children with cerebral palsy. *IEEE Access*, 11:65906–65923, 2023.

[45] Cheng Song, Lu Lu, Zhen Ke, Long Gao, and Shuai Ding. Self-supervised gait-based emotion representation learning from selective strongly augmented skeleton sequences, 2024.

[46] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1987.

[47] P. Srihari and Jonnadula Harikiran. Skeleton based human activity prediction in gait thermal images using siamese networks. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pages 1163–1170, 2022.

[48] Jan Stenum, Melody M. Hsu, Alexander Y. Pantelyat, and Ryan T. Roemmich. Clinical gait analysis using video-based pose estimation: Multiple perspectives, clinical populations, and measuring change. *PLOS Digital Health*, 3(3):1–23, 03 2024.

[49] Guangmin Sun and Zhongqi Wang. Fall detection algorithm for the elderly based on human posture estimation. In *2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pages 172–176, 2020.

[50] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition, 2022.

[51] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[52] Chi Xu, Yasushi Makihara, Xiang Li, and Yasushi Yagi. Occlusion-aware human mesh model-based gait recognition. *IEEE transactions on information forensics and security*, 18:1309–1321, 2023.

[53] Shihao Xu, Jing Fang, Xiping Hu, Edith Ngai, Wei Wang, Yi Guo, and Victor C. M. Leung. Emotion recognition from gait analyses: Current research and future directions. *IEEE Transactions on Computational Social Systems*, pages 1–15, 2022.

[54] Yu-Hung Yeh, Jiun-Lin Yan, Meng-Xun Gu, Yi-Wei Chen, and Ta-Sung Lee. Frequency-domain analysis for accurate and robust gait cycle time detection with clinical data. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4200–4204, 2022.

[55] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444, 2006.

Appendix A

Table 5.1: Accuracy of the matching by using joint angles as a feature on NM sub-dataset. These results are from employing 2D joints extracted by MediaPipe.

	Whole			Upper			Lower		
	20	49	118	20	49	118	20	49	118
0°	0.1	0.06	0.10	0.3	0.24	0.19	0.1	0.06	0.05
18°	0.1	0.12	0.17	0.25	0.20	0.20	0.1	0.08	0.14
36°	0.2	0.22	0.18	0.25	0.33	0.20	0.25	0.22	0.18
54°	0.4	0.24	0.25	0.35	0.37	0.31	0.35	0.12	0.14
72°	0.45	0.31	0.20	0.55	0.35	0.29	0.25	0.12	0.16
90°	0.45	0.31	0.25	0.25	0.24	0.19	0.35	0.22	0.20
108°	0.35	0.27	0.25	0.3	0.20	0.14	0.4	0.27	0.20
126°	0.5	0.33	0.29	0.4	0.27	0.21	0.45	0.27	0.22
144°	0.5	0.45	0.31	0.45	0.27	0.15	0.45	0.27	0.24
162°	0.55	0.39	0.26	0.45	0.24	0.12	0.6	0.31	0.25
180°	0.45	0.16	0.14	0.4	0.20	0.09	0.25	0.12	0.09
Max	0.55	0.45	0.31	0.55	0.37	0.31	0.6	0.31	0.25
Voting	0.6	0.55	0.48	0.65	0.59	0.48	0.5	0.41	0.4

Table 5.2: Accuracy of the matching by using time-dependent correlation as a feature on NM sub-dataset. These results are from employing 2D joints extracted by MediaPipe.

	Whole			Upper			Lower		
	20	49	118	20	49	118	20	49	118
0°	0.15	0.14	0.14	0.00	0.00	0.00	0.10	0.10	0.09
18°	0.25	0.20	0.17	0.00	0.00	0.00	0.20	0.10	0.08
36°	0.35	0.33	0.21	0.00	0.00	0.00	0.20	0.20	0.10
54°	0.25	0.16	0.14	0.00	0.00	0.00	0.15	0.08	0.08
72°	0.30	0.20	0.08	0.00	0.00	0.00	0.25	0.10	0.08
90°	0.15	0.16	0.08	0.00	0.00	0.00	0.30	0.20	0.08
108°	0.15	0.14	0.09	0.00	0.00	0.00	0.10	0.06	0.03
126°	0.10	0.08	0.09	0.00	0.00	0.00	0.15	0.08	0.09
144°	0.20	0.14	0.07	0.00	0.00	0.00	0.20	0.10	0.05
162°	0.30	0.18	0.07	0.00	0.00	0.00	0.30	0.12	0.07
180°	0.20	0.08	0.03	0.00	0.00	0.00	0.20	0.04	0.01
Max	0.35	0.33	0.21	0.00	0.00	0.00	0.30	0.20	0.10
Voting	0.35	0.37	0.28	0.00	0.00	0.00	0.30	0.27	0.20

Table 5.3: Accuracy of the matching by using joint angles and time-dependent correlation as features on NM sub-dataset. These results are from employing 2D joints extracted by MediaPipe.

	Whole			Upper			Lower		
	20	49	118	20	49	118	20	49	118
0°	0.05	0.08	0.01	-	-	-	0.00	0.06	0.00
18°	0.20	0.06	0.04	-	-	-	0.15	0.06	0.03
36°	0.00	0.02	0.04	-	-	-	0.00	0.00	0.03
54°	0.00	0.06	0.02	-	-	-	0.00	0.00	0.02
72°	0.30	0.10	0.05	-	-	-	0.30	0.06	0.03
90°	0.10	0.00	0.03	-	-	-	0.05	0.00	0.03
108°	0.00	0.06	0.01	-	-	-	0.10	0.04	0.01
126°	0.10	0.04	0.03	-	-	-	0.10	0.06	0.02
144°	0.10	0.04	0.01	-	-	-	0.10	0.06	0.03
162°	0.20	0.04	0.03	-	-	-	0.20	0.08	0.02
180°	0.20	0.04	0.00	-	-	-	0.00	0.04	0.00
Max	0.30	0.10	0.05	-	-	-	0.30	0.08	0.03
Voting	0.15	0.12	0.03	-	-	-	0.15	0.08	0.03

Table 5.4: Accuracy of the matching by using time-independent correlation as a feature on NM sub-dataset. These results are from employing 2D joints extracted by MediaPipe.

	Whole			Upper			Lower		
	20	49	118	20	49	118	20	49	118
0°	0.4	0.43	0.36	0.05	0.06	0.03	0.25	0.20	0.19
18°	0.4	0.29	0.20	0.1	0.00	0.00	0.25	0.20	0.12
36°	0.6	0.37	0.35	0.15	0.02	0.01	0.4	0.20	0.23
54°	0.65	0.45	0.32	0.1	0.02	0.01	0.5	0.37	0.24
72°	0.55	0.33	0.23	0.15	0.02	0.03	0.3	0.20	0.14
90°	0.55	0.47	0.35	0.1	0.02	0.03	0.4	0.31	0.21
108°	0.45	0.35	0.27	0.15	0.04	0.02	0.25	0.16	0.15
126°	0.55	0.53	0.27	0.1	0.06	0.03	0.45	0.33	0.25
144°	0.75	0.59	0.46	0.05	0.04	0.01	0.6	0.41	0.30
162°	0.5	0.39	0.23	0.15	0.04	0.02	0.4	0.27	0.20
180°	0.4	0.27	0.08	0.2	0.04	0.02	0.35	0.20	0.11
Max	0.75	0.59	0.46	0.2	0.06	0.03	0.6	0.41	0.30
Voting	0.9	0.88	0.74	0.3	0.04	0.01	0.75	0.63	0.59

Table 5.5: Accuracy of the matching by using joint angles as a feature on NM sub-dataset. These results are from employing 3D joints extracted by MediaPipe.

	Whole			Upper			Lower		
	20	49	118	20	49	118	20	49	118
0°	0.60	0.47	0.34	0.45	0.24	0.12	0.50	0.35	0.29
18°	0.40	0.33	0.28	0.10	0.12	0.14	0.35	0.33	0.25
36°	0.45	0.35	0.30	0.20	0.14	0.10	0.40	0.37	0.30
54°	0.55	0.39	0.40	0.35	0.35	0.22	0.60	0.39	0.34
72°	0.55	0.39	0.37	0.55	0.35	0.24	0.50	0.41	0.33
90°	0.50	0.47	0.37	0.25	0.20	0.22	0.55	0.37	0.31
108°	0.60	0.37	0.37	0.25	0.20	0.24	0.60	0.37	0.31
126°	0.45	0.51	0.44	0.30	0.31	0.16	0.40	0.47	0.36
144°	0.55	0.45	0.42	0.35	0.31	0.17	0.55	0.39	0.36
162°	0.65	0.49	0.36	0.35	0.22	0.12	0.65	0.51	0.36
180°	0.65	0.47	0.31	0.30	0.16	0.13	0.65	0.45	0.29
Max	0.65	0.51	0.44	0.55	0.35	0.24	0.65	0.51	0.36
Voting	0.80	0.67	0.66	0.50	0.51	0.41	0.75	0.71	0.59

Table 5.6: Accuracy of the matching by using time-independent correlation as a feature on NM sub-dataset. These results are from employing 3D joints extracted by MediaPipe.

	Whole			Upper			Lower		
	20	49	118	20	49	118	20	49	118
0°	0.35	0.29	0.23	0.00	0.06	0.02	0.50	0.24	0.16
18°	0.25	0.27	0.16	0.05	0.00	0.01	0.25	0.29	0.16
36°	0.50	0.39	0.25	0.20	0.02	0.00	0.50	0.31	0.14
54°	0.40	0.24	0.21	0.15	0.02	0.01	0.30	0.18	0.11
72°	0.55	0.33	0.16	0.15	0.12	0.05	0.45	0.27	0.09
90°	0.35	0.27	0.19	0.15	0.04	0.04	0.15	0.10	0.08
108°	0.35	0.35	0.23	0.10	0.04	0.03	0.25	0.18	0.14
126°	0.60	0.27	0.24	0.10	0.04	0.02	0.40	0.20	0.17
144°	0.40	0.35	0.35	0.10	0.08	0.03	0.30	0.24	0.25
162°	0.60	0.39	0.26	0.30	0.02	0.00	0.50	0.37	0.26
180°	0.50	0.27	0.14	0.10	0.06	0.01	0.40	0.16	0.08
Max	0.60	0.39	0.35	0.30	0.12	0.05	0.50	0.37	0.26
Voting	0.85	0.74	0.67	0.20	0.06	0.04	0.85	0.63	0.50

Table 5.7: Accuracy of the matching by using joint angles as a feature on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by OpenPose.

	Whole			Upper			Lower		
	20	50	100	20	50	100	20	50	100
0°	0.35	0.14	0.12	0.25	0.10	0.08	0.30	0.12	0.07
15°	0.40	0.24	0.22	0.10	0.14	0.08	0.25	0.18	0.16
30°	0.25	0.24	0.20	0.30	0.18	0.14	0.30	0.28	0.21
45°	0.30	0.12	0.09	0.40	0.16	0.09	0.20	0.10	0.12
60°	0.20	0.14	0.12	0.15	0.08	0.04	0.20	0.20	0.16
75°	0.30	0.22	0.11	0.10	0.10	0.06	0.30	0.36	0.28
90°	0.30	0.16	0.15	0.20	0.16	0.11	0.30	0.26	0.21
180°	0.30	0.22	0.18	0.25	0.16	0.13	0.20	0.18	0.11
195°	0.20	0.12	0.11	0.20	0.14	0.09	0.25	0.20	0.19
210°	0.40	0.20	0.13	0.30	0.14	0.06	0.25	0.24	0.14
225°	0.10	0.04	0.07	0.05	0.00	0.06	0.20	0.14	0.17
240°	0.20	0.10	0.11	0.15	0.08	0.07	0.35	0.14	0.15
255°	0.20	0.04	0.07	0.15	0.00	0.05	0.35	0.16	0.15
270°	0.30	0.16	0.12	0.15	0.06	0.05	0.40	0.20	0.17
Max	0.40	0.24	0.22	0.40	0.18	0.14	0.40	0.36	0.28
Voting	0.65	0.42	0.36	0.55	0.30	0.21	0.55	0.52	0.47

Table 5.8: Accuracy of the matching by using joint angles as a feature on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by AlphaPose.

	Whole			Upper			Lower		
	20	50	100	20	50	100	20	50	100
0°	0.55	0.26	0.12	0.30	0.12	0.08	0.45	0.26	0.07
15°	0.40	0.34	0.22	0.25	0.14	0.08	0.30	0.28	0.16
30°	0.40	0.22	0.20	0.25	0.14	0.14	0.25	0.22	0.21
45°	0.30	0.22	0.09	0.25	0.20	0.09	0.20	0.16	0.12
60°	0.20	0.24	0.12	0.20	0.20	0.04	0.20	0.24	0.16
75°	0.25	0.32	0.11	0.20	0.22	0.06	0.20	0.32	0.28
90°	0.25	0.32	0.15	0.25	0.22	0.11	0.30	0.24	0.21
180°	0.30	0.26	0.18	0.25	0.20	0.13	0.15	0.20	0.11
195°	0.30	0.18	0.11	0.20	0.14	0.09	0.30	0.20	0.19
210°	0.40	0.20	0.13	0.25	0.14	0.06	0.35	0.22	0.14
225°	0.25	0.16	0.07	0.35	0.14	0.06	0.30	0.16	0.17
240°	0.30	0.20	0.11	0.35	0.10	0.07	0.30	0.18	0.15
255°	0.35	0.14	0.07	0.30	0.12	0.05	0.45	0.16	0.15
270°	0.25	0.16	0.12	0.25	0.08	0.05	0.20	0.10	0.17
Max	0.55	0.34	0.22	0.35	0.22	0.14	0.45	0.32	0.28
Voting	0.65	0.48	0.53	0.50	0.46	0.21	0.60	0.54	0.47

Table 5.9: Accuracy of the matching by using time-dependent correlation as a feature on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by OpenPose.

	Whole			Upper			Lower		
	20	50	100	20	50	100	20	50	100
0°	0.15	0.06	0.03	0.00	0.00	0.00	0.15	0.00	0.00
15°	0.05	0.06	0.02	0.00	0.00	0.00	0.00	0.00	0.00
30°	0.10	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
45°	0.15	0.06	0.01	0.00	0.00	0.00	0.05	0.04	0.02
60°	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.01
75°	0.05	0.02	0.04	0.00	0.00	0.00	0.00	0.00	0.01
90°	0.05	0.04	0.05	0.00	0.00	0.00	0.05	0.02	0.01
180°	0.10	0.08	0.01	0.00	0.00	0.00	0.05	0.02	0.01
195°	0.15	0.06	0.05	0.00	0.00	0.00	0.15	0.04	0.02
210°	0.15	0.04	0.02	0.00	0.00	0.00	0.00	0.02	0.01
225°	0.25	0.08	0.05	0.00	0.00	0.00	0.05	0.02	0.01
240°	0.30	0.10	0.06	0.00	0.00	0.00	0.05	0.04	0.03
255°	0.10	0.04	0.02	0.00	0.00	0.00	0.05	0.02	0.01
270°	0.25	0.10	0.04	0.00	0.00	0.00	0.10	0.00	0.00
Max	0.30	0.10	0.06	0.00	0.00	0.00	0.15	0.04	0.03
Voting	0.20	0.14	0.08	0.00	0.00	0.00	0.00	0.00	0.00

Table 5.10: Accuracy of the matching by using time-dependent correlation as a feature on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by AlphaPose.

	Whole			Upper			Lower		
	20	50	100	20	50	100	20	50	100
0°	0.05	0.02	0.02	0.00	0.00	0.00	0.05	0.00	0.00
15°	0.20	0.06	0.04	0.00	0.00	0.00	0.15	0.06	0.03
30°	0.20	0.08	0.06	0.00	0.00	0.00	0.05	0.02	0.00
45°	0.15	0.14	0.10	0.00	0.00	0.00	0.05	0.02	0.01
60°	0.25	0.12	0.08	0.00	0.00	0.00	0.10	0.04	0.03
75°	0.15	0.06	0.03	0.00	0.00	0.00	0.05	0.02	0.01
90°	0.05	0.12	0.07	0.00	0.00	0.00	0.00	0.00	0.00
180°	0.05	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
195°	0.15	0.12	0.07	0.00	0.00	0.00	0.05	0.02	0.01
210°	0.35	0.18	0.09	0.00	0.00	0.00	0.15	0.06	0.05
225°	0.15	0.10	0.06	0.00	0.00	0.00	0.10	0.02	0.02
240°	0.10	0.04	0.04	0.00	0.00	0.00	0.15	0.04	0.02
255°	0.10	0.04	0.06	0.00	0.00	0.00	0.00	0.00	0.00
270°	0.05	0.04	0.03	0.00	0.00	0.00	0.10	0.00	0.00
Max	0.35	0.18	0.10	0.00	0.00	0.00	0.15	0.06	0.05
Voting	0.30	0.18	0.13	0.00	0.00	0.00	0.15	0.04	0.01

Table 5.11: Accuracy of the matching by using joint angles and time-dependent correlation as features on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by OpenPose.

	Whole			Upper			Lower		
	20	50	100	20	50	100	20	50	100
0°	0.25	0.12	0.04	-	-	-	0.20	0.04	0.02
15°	0.25	0.08	0.05	-	-	-	0.10	0.06	0.01
30°	0.15	0.08	0.08	-	-	-	0.15	0.08	0.05
45°	0.05	0.02	0.01	-	-	-	0.05	0.08	0.03
60°	0.05	0.04	0.04	-	-	-	0.10	0.06	0.03
75°	0.20	0.10	0.04	-	-	-	0.00	0.06	0.03
90°	0.05	0.10	0.10	-	-	-	0.05	0.12	0.08
180°	0.10	0.06	0.04	-	-	-	0.05	0.00	0.00
195°	0.00	0.08	0.02	-	-	-	0.00	0.06	0.04
210°	0.00	0.04	0.02	-	-	-	0.10	0.08	0.06
225°	0.10	0.00	0.02	-	-	-	0.10	0.02	0.04
240°	0.15	0.06	0.06	-	-	-	0.10	0.02	0.05
255°	0.10	0.04	0.03	-	-	-	0.10	0.02	0.02
270°	0.20	0.06	0.04	-	-	-	0.05	0.04	0.04
Max	0.25	0.12	0.10	-	-	-	0.20	0.12	0.08
Voting	0.25	0.18	0.14	-	-	-	0.15	0.14	0.11

Table 5.12: Accuracy of the matching by using joint angles and time-dependent correlation as features on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by AlphaPose.

	Whole			Upper			Lower		
	20	50	100	20	50	100	20	50	100
0°	0.15	0.04	0.03	-	-	-	0.05	0.04	0.02
15°	0.20	0.08	0.04	-	-	-	0.15	0.04	0.01
30°	0.20	0.08	0.03	-	-	-	0.15	0.08	0.03
45°	0.05	0.08	0.06	-	-	-	0.05	0.04	0.05
60°	0.20	0.10	0.06	-	-	-	0.20	0.16	0.05
75°	0.05	0.12	0.05	-	-	-	0.10	0.04	0.04
90°	0.20	0.12	0.08	-	-	-	0.15	0.08	0.08
180°	0.05	0.06	0.03	-	-	-	0.05	0.04	0.06
195°	0.10	0.08	0.04	-	-	-	0.10	0.06	0.05
210°	0.30	0.12	0.05	-	-	-	0.25	0.08	0.01
225°	0.10	0.06	0.05	-	-	-	0.15	0.10	0.02
240°	0.15	0.04	0.06	-	-	-	0.05	0.04	0.03
255°	0.20	0.10	0.06	-	-	-	0.00	0.04	0.02
270°	0.10	0.08	0.06	-	-	-	0.20	0.08	0.04
Max	0.30	0.12	0.08	-	-	-	0.25	0.16	0.08
Voting	0.35	0.22	0.15	-	-	-	0.10	0.14	0.08

Table 5.13: Accuracy of the matching by using time-independent correlation as a feature on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by OpenPose.

	Whole			Upper			Lower		
	20	50	100	20	50	100	20	50	100
0°	0.15	0.10	0.07	0.00	0.00	0.01	0.10	0.02	0.02
15°	0.30	0.10	0.10	0.10	0.00	0.00	0.10	0.08	0.04
30°	0.25	0.18	0.16	0.15	0.02	0.01	0.10	0.04	0.07
45°	0.45	0.18	0.16	0.05	0.02	0.01	0.25	0.08	0.06
60°	0.40	0.30	0.25	0.15	0.10	0.05	0.20	0.10	0.11
75°	0.40	0.22	0.19	0.05	0.04	0.02	0.30	0.18	0.12
90°	0.35	0.24	0.16	0.05	0.02	0.01	0.40	0.14	0.05
180°	0.20	0.18	0.09	0.15	0.08	0.03	0.15	0.04	0.04
195°	0.30	0.12	0.07	0.10	0.06	0.03	0.15	0.08	0.00
210°	0.45	0.30	0.15	0.25	0.08	0.01	0.20	0.06	0.04
225°	0.60	0.28	0.20	0.15	0.10	0.04	0.40	0.18	0.08
240°	0.40	0.28	0.22	0.10	0.02	0.01	0.35	0.16	0.11
255°	0.35	0.28	0.18	0.10	0.02	0.00	0.25	0.18	0.08
270°	0.50	0.28	0.13	0.00	0.00	0.00	0.35	0.18	0.10
Max	0.60	0.30	0.25	0.25	0.10	0.05	0.40	0.18	0.12
Voting	0.80	0.64	0.47	0.20	0.04	0.01	0.60	0.34	0.17

Table 5.14: Accuracy of the matching by using time-independent correlation as a feature on OUMVLP-Pose dataset. These results are from employing 2D joints extracted by AlphaPose.

	Whole			Upper			Lower		
	20	50	100	20	50	100	20	50	100
0°	0.35	0.26	0.13	0.15	0.02	0.00	0.40	0.10	0.09
15°	0.55	0.38	0.20	0.05	0.02	0.00	0.50	0.16	0.13
30°	0.30	0.24	0.21	0.15	0.04	0.01	0.35	0.16	0.13
45°	0.40	0.22	0.10	0.20	0.06	0.02	0.45	0.20	0.10
60°	0.40	0.38	0.31	0.10	0.06	0.02	0.50	0.18	0.18
75°	0.25	0.24	0.23	0.00	0.00	0.00	0.35	0.30	0.16
90°	0.50	0.34	0.24	0.10	0.04	0.02	0.30	0.20	0.12
180°	0.30	0.20	0.17	0.15	0.04	0.03	0.15	0.10	0.07
195°	0.35	0.20	0.11	0.15	0.10	0.04	0.25	0.12	0.05
210°	0.30	0.20	0.14	0.10	0.00	0.01	0.25	0.18	0.07
225°	0.60	0.36	0.25	0.05	0.02	0.00	0.35	0.16	0.09
240°	0.45	0.30	0.20	0.20	0.02	0.02	0.35	0.26	0.15
255°	0.50	0.20	0.20	0.00	0.02	0.01	0.25	0.12	0.11
270°	0.35	0.26	0.21	0.05	0.00	0.00	0.25	0.10	0.08
Max	0.60	0.38	0.31	0.20	0.10	0.04	0.50	0.30	0.18
Voting	0.75	0.70	0.60	0.25	0.02	0.01	0.60	0.44	0.33