ACKNOWLEDGEMENT

Upon the completion of this report, there are many works involved that need the help of various people. They have guided me thoroughly upon understanding the concept of the project, the essence of knowledge provided will not be at waste. Particularly, I would like to express my biggest thanks to my main project supervisor, Miss Nur Maisarah Binti Mohd Sobran of her guidance for the project, suggesting ideas, help in understanding project purpose and concept and thoroughly guide me upon this whole project. Without her guidance, the project might unable to be implementing well and may not be able to succeed.

I am also very thankful to UTeM administration especially FKE administration for the funding of material of the project, helping much in implementing the hardware part of the project.

Besides that, thanks to my fellow course mates that help giving friendly advice, suggestions, for me to complete the system well and help me on understanding some of the confusing part of the project.

ABSTRACT

Along the decade, researchers have proposed some various methods to represents the human walking behaviors in either individual or in crowd environment. Human walking behaviors can be divided into several viewpoints such as route selection, navigation, and path finding. Although human walking behavior is unpredictable and has dynamic characteristics, these viewpoints are the most dominant behavior that human always considered when they are walking. Furthermore, many researchers have widen their scope of investigation as they are more focusing in human walking behaviors under panic situations that usually in crowded conditions due to natural disaster or terrorist attacks. In this project, a method to analyze human walking data is proposed and human walking data will be analysed by classifying the path taken by human while walking whether right, left of a forward path. However in order to collect human walking data, an experiment is set up which involves five different subjects with different gender, height and weight considered. In the experiment, the subject will wear a wearable device containing inertial measurement unit around their waist and walks along a created path. Inertial measurement unit is used to determine the orientation and position of human while they are walking. The data considered from inertial measurement unit is Yaw, Pitch, and Roll data. The human walking data will be classified by using unsupervised learning method because human walking data are unclassified data and the results of the analysis are cannot be predicted. K-means clustering method is used to classified human walking activity. The number of cluster will be determined by using K-means clustering. After that, the performance measurement of K-means clustering is carried out to evaluate the performance of K-means clustering. Silhouette Coefficient method is used for this purpose. The validity of number of cluster for clustering human walking activity is determined. As conclusion, the number of cluster that is suitable to classified human walking data is three.

ABSTRAK

Sepanjang dekad ini, penyelidik telah mencadangkan beberapa pelbagai kaedah untuk mewakili tingkah laku berjalan kaki manusia sama ada individu atau dalam persekitaran orang ramai. Tingkah laku yang berjalan kaki manusia boleh dibahagikan kepada beberapa pandangan seperti pemilihan laluan, navigasi, dan laluan dapatan. Walaupun tingkah laku berjalan manusia adalah tidak menentu dan mempunyai ciri-ciri dinamik, pandangan ini adalah tingkah laku yang paling dominan manusia yang sentiasa dipertimbangkan apabila mereka berjalan. Tambahan pula, ramai penyelidik telah memperluaskan skop siasatan kerana mereka lebih memberi tumpuan dalam tingkah laku berjalan kaki manusia di bawah keadaan panik yang biasanya dalam keadaan sesak disebabkan oleh bencana alam atau serangan pengganas. Dalam projek ini, kaedah untuk menklasifikasikan data manusia berjalan telah dicadangkan dan data berjalan kaki manusia akan dianalisis dengan mengklasifikasikan jalan yang diambil oleh manusia ketika berjalan sama ada kanan, kiri jalan ke hadapan. Walau bagaimanapun dalam usaha untuk mengumpul data berjalan kaki manusia, eksperimen ditubuhkan yang melibatkan lima mata subjek yang berbeza dengan jantina berbeza, ketinggian dan berat badan dipertimbangkan. Dalam eksperimen ini, subjek menggunakan peranti yang mengandungi unit pengukuran inersia di pinggang mereka dan berjalan di sepanjang jalan yang diwujudkan. Unit pengukuran inersia digunakan untuk menentukan orientasi dan kedudukan manusia sementara mereka berjalan. Data yang berjalan kaki manusia akan diklasifikasikan dengan menggunakan kaedah pembelajaran tanpa pengawasan kerana data berjalan manusia adalah data tidak dikelaskan dan keputusan analisis tidak dapat diramalkan. cara kaedah pengelompokan K digunakan untuk aktiviti berjalan kaki manusia terperingkat. Bilangan kluster akan ditentukan dengan menggunakan cara berkelompok K. Selepas itu, pengukuran prestasi K-cara kelompok dijalankan untuk menilai prestasi kelompok K. Kaedah Pekali Bayang digunakan untuk tujuan ini. Kesahihan beberapa kelompok untuk kelompok aktiviti berjalan kaki manusia ditentukan. Kesimpulannya, bilangan kelompok yang sesuai untuk terperingkat data berjalan manusia adalah tiga.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The main goal of this project is to develop a method to classify an unclassified human walking data by using unsupervised learning and to developed experimental setups to get the human walking data. This chapter will cover the research background, motivation, problem statement, objectives and scope of the research. A further explanation also will be included here to provide understanding for the user about this project.

1.2 Project Research Background

Along the decade, researchers have proposed some various methods to represents the human walking behaviors in either individual or in crowd environment. Human walking behaviors can be divided into several viewpoints such as route selection, navigation, and path finding [11]. Although human walking behavior is unpredictable and has dynamic characteristics, these viewpoints are the most dominant behavior that human always considered when they are walking [6]. Furthermore, many researchers have widen their scope of investigation as they are more focusing in human walking behaviors under panic situations that usually in crowded conditions due to natural disaster or terrorist attacks [7-9]. Zheng et al. [7], has proposed seven methods for crowd evacuations that also includes the important aspects for human walking behaviors which cover heterogeneity, scale of modeling, condition, space, and time steps.

In this project, human walking data will be analysed by classifying the path taken by human while walking whether right, left of a forward path. The human walking data will be classified by using unsupervised learning method because human walking data are unclassified data and the results of the analysis are cannot be predicted. Therefore, unsupervised learning method is very suitable as it is used to clustering unclassified data and the results of clustering are varies according to the users. Moreover, it is used for large number of dataset as human walking data consist of a large number of data.

1.3 Motivation

Knowledge Discovery in Database can be defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. It is an interactive and iterative process involving several steps such as selection, pre-processing, transformation, data mining and interpretation of data. Knowledge in data is very important as the knowledge will help to identify and interpret the data which can be used later in the future.

As reported by CIO Magazine's cover story in May 1998 [1], about 13 million customers will contact the customer service call centre of Bank of America in order to listen to the marketing advertisement. Therefore, Bank of America conclude that the costumer like to be informed of new product or services. So in this case, knowledge discovery in database is important as they help to identify the type of marketing approach for certain customer based on their profile.

Besides that, nearly 3000 cases per year involving brain tumors among children have been reported in United States. Among all the cases, nearly half of them are considered fatal. According to Director of brain tumor research at Children's Memorial Hospital in Chicago; Eric Bremer [2], they have set a goal to create a database of gene expression of the patients, in order to give them a better treatment. Therefore, they used software of data mining which is Clementine data mining software; developed by SPSS, Inc. to classify the types of tumor into 12 or so cluster types.

In November 2002, former President of United States of America, President Bill Clinton spoke at Democratic Leadership Council [3], mentioned that after the event of 11 September 2001, the Federal Bureau of Investigation (FBI) has received a great amount of data considering five terrorist related to the incident. As stated from the data, one of the terrorist possessed 30 credit cards with some combined balances and stay in the country for almost 2 years. Moreover, President Bill Clinton concluded that further investigations are needed to gain some information for the data.

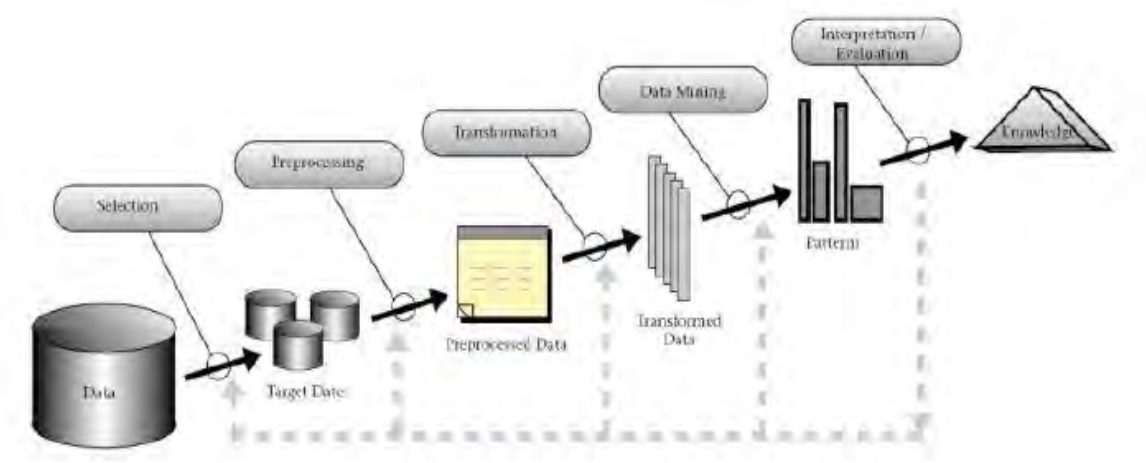


Figure 1.1: Steps in Knowledge Discovering of Database [4].

1.4 Problem Statements

In achieving this project, there are several problems that have been encountered. First of all, the problems face is to select the most suitable clustering method that need to be used to classify the human walking data. The method used need to accurately classify the human walking data as many data are involved and the data have different range of value. Therefore, unsuitable clustering method will result inaccurately classification of data.

Next, the experimental setup to collecting the human walking data is needs to take serious part in succeeding the project. The experimental setup has able to collect the position and orientation of human while walking. Moreover, every aspect for the experiments such as the sensor used and the walking path have to be considered. The sensor need to detect and track the position which means it has to accurately identify the position and orientation of human. Therefore, a single sensor cannot achieve this accurate data and sensor fusion process has to take place. The sensor fusion also can correct the deficiencies of the single sensor.

The height of the human must also be considered in running this project. The sensor will be placed on the human body at their waist in order to examine the positon and orientation of human. Different height of human will result the different height position of waist of a person. If the person is taller, this means that the angle of elevation will be different from a short person as higher position will has a bigger angle of elevation.

1.5 Objectives

The objectives of this project are:

1) To develop an experimental setup to collect human walking data.
2) To develop a method to classify an unclassified human walking data by using unsupervised learning.
3) To evaluate the performance of the k-means clustering by using Silhoutte Coefficient method.

1.6 Scope

The scope below should be emphasized in order to achieve the objectives of the project.

1) The sensors have to be able to identify the position and orientation of human whenever in static or walking motions. This means that the sensors have gyroscopic features and can compute data from X-Y-Z plane.
2) The method must be able to classify human walking data as there are many data produced during the experiment.
3) The performance of k-means clustering will be discussed in term of number of suitable k clusters.

CHAPTER 2

PROJECT BACKGROUND & LITERATURE REVIEW

2.1 Introduction

In this chapter, the review of related previous research project will be discussed. The information obtained will become some useful source that can be used as references in order to make this project successful. Other than that, the obtained information will be synthesized to some literature review by integrating the information to evaluate them. Therefore, this chapter will discuss about the literature review.

2.2 Project Background

Machine learning technique is very popular technique nowadays as it is used to train computer to explore and study the algorithm that can be learn and interpret some data. Machine learning will gives computer the ability to learn and improving performance without being programmed [5]. There are two types of machine learning; supervised learning and unsupervised learning. These two types have different functions and different characteristics which can be applied according to certain condition.

2.2.1 Supervised Learning

In supervised learning, the training data will includes both the input and the desired result. In other words, the correct results or target are known first and the input to the model are given during the learning process. The supervised learning will find a way to build a model

that can predict the response value of dataset. This method is popular because it gives fast and accurate results. Moreover, supervised learning are able to generalize the data which means that it is able to give reasonable outputs even new inputs are given without knowing a proper target. There are two categories of algorithm in supervised learning which are classification and regression. Classification is used to assign membership of group for data samples. There are two ways to assign new values to given class. Firsts by using crispy classification method; which the classifier will return the input label. Seconds by using probabilistic classification method; which the classifier will return the input probabilities to the belonging class.

Moreover, there are three tools of supervised learning; neural networks, support vector machine, and decision tree [6]. Neural Networks refer to Artificial Neural Network (ANN) that consists of a set of interconnected simple processing units (neurons or nodes) which combine to output a signal to solve a certain problem based on the input signals it received [7]. The neural network also inspired by human brain which contains highly amount of connected neurons. Figure below shows the graphical structure of neural network.

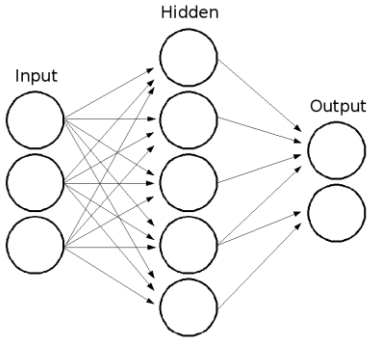


Figure 2.1: Structure of Neural Networks [4].

A Neural Network is usually contains three layers which is an input layer, a hidden layer, and an output layer. The main property of neural networks is that it has adjustable gains that are slowly adjusted through iterations influenced by the input-output patterns.

Support vector machine (SVM) is a method used for classification, regression and outlier detection. It means that the data will be analyzed and the pattern of data will be recognized by mapping vectors of input feature to higher dimensional space [7]. This method

uses subset of training to make decision; therefore it is also a memory efficient. A standard support vector machine is non-probabilistic linear classifier that categorizes the new input. A good separation is achieved if the hyperplane which has largest distance is nearer to training point to minimize the generalization error [7]. Figure 2.2 shows the good separation of data by using support vector machine method.

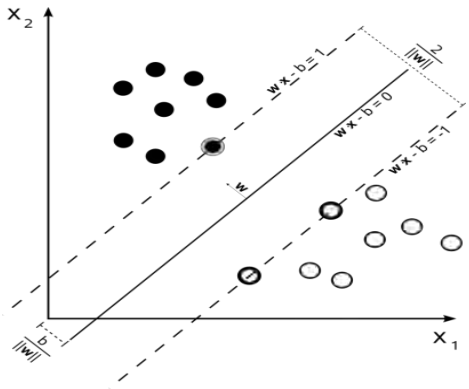


Figure 2.2: A good separation of data [7].

Another unsupervised learning tool is decision tree method. Decision tree method is a tree-like model used to approximating discrete-valued function that is robust to noisy data and able to learn functions. This method is very popular as is has been successfully applied to broad range of task from learning to diagnose medical cases. As an example, the decision tree learning been applied to classify medical patients by their disease and equipment malfunction by their cause which generally the task is to classify the data into a set of possible categories. Figure 2.3 shows the example of decision tree method.

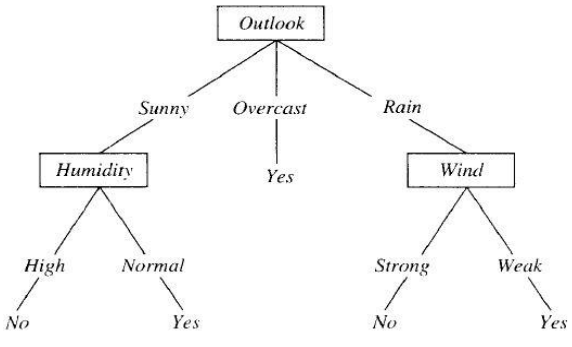


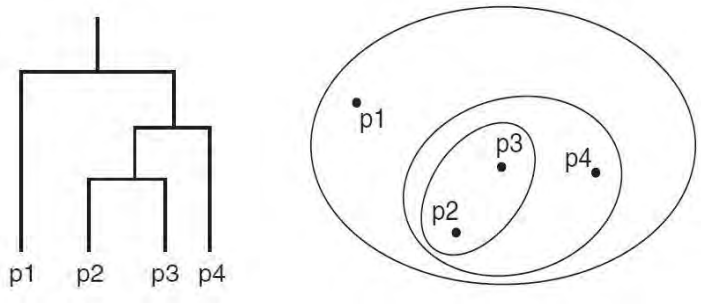
Figure 2.3: Examples of decision tree concept [8]

2.2.2 Unsupervised Learning

In unsupervised learning, the training data is not provided with the predicted results during the training. In other words, the results of data of unsupervised learning will vary according to user need. This means that there will no incorrect results when unsupervised learning method is used. This method is different from supervised learning method as it is not allow for prediction of result compared to supervised data which the result for the training is provided and can be predicted [6]. Moreover, unsupervised learning can be used to cluster the input data according to classes based on their statistical properties only. Even though this method does not require labels for clustering, the labelling can be carried out even if they are only available for small number of objects in desired class.

Clustering can be defined as a process of dividing a set of data or objects into a set of sub-classes, called clusters. There are two types of clustering in unsupervised learning; hierarchical clustering and partitional clustering [6]. A hierarchical clustering is a set of nested sub-classes or clusters that are organized as a tree which each cluster in the tree is the combination of its sub-clusters, and the root is cluster containing all objects [9].

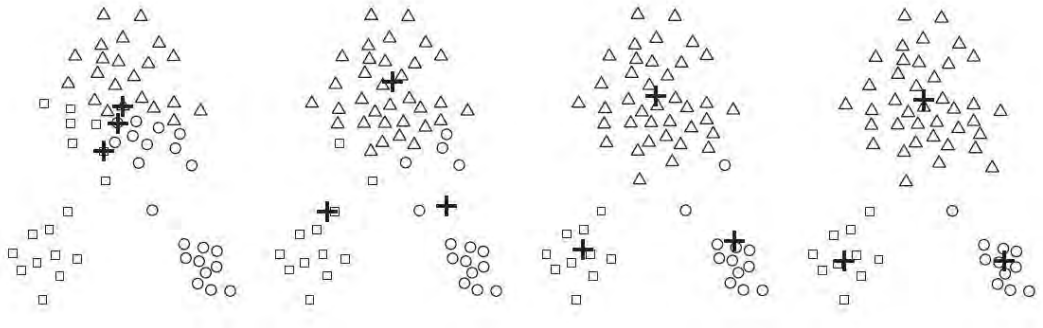
In hierarchical clustering, it will always finds successive cluster by previously refers to actualized cluster. Hierarchical clustering is generated based on two basic approaches; agglomerative and divisive. The difference between agglomerative hierarchical clustering technique and divisive hierarchical clustering technique is that the agglomerative will start with points of each individual element in separate cluster merging with closest cluster pairs while divisive start with single cluster which then be splatted until clusters of individual points are remain. Figure 2.4 shows the graphical representation of hierarchical clustering.



(a) Dendrogram (b) Nested cluster diagram

Figure 2.4: Two types of hierarchical clustering representation [9].

Partitional clustering is referring to a subset which contains each division of set of data objects that transformed into non-overlapping clusters. This clustering method usually determines all clusters at ones. There are two approaches in generating partitional clustering; K-means clustering and DBSCAN clustering [9]. K-means clustering is a clustering method which separates the data into number of k clusters, based on their characteristics. Each cluster will be represented by the centroid; center of cluster and each nearest point to the centroid will be assigned in that particular cluster. The main purpose of K-means clustering method is to minimize the distance between cluster centroids with corresponding data. Figure 2.5 below shows the examples of k-means clustering with four number of iteration.



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Figure 2.5: Examples of K-means Clustering [9].

DBSCAN or density-based clustering algorithm is a partitional clustering that automatically determines the number of cluster of data. However, this clustering method is

considered as incomplete clustering because it only considered points in high and medium density only as data while points in low-density are considered as noise and omitted.

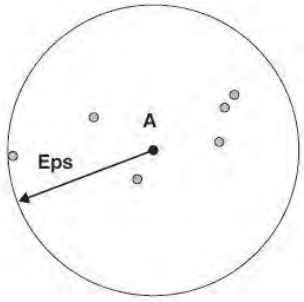


Figure 2.6: DBSCAN clustering [9].

2.3 Differences between Supervised Learning and Unsupervised Learning

Table 2.1: Difference between Supervised and Unsupervised.

Supervised Learning	Unsupervised Learning
1. Training data includes both the input and the desired result during the training.	1. Training data is not provided with the predicted results during the training.
2. Find a way to build a model that can predict the response value of dataset.	2. Not allow for prediction of result.
3. Able to generalize the data even new inputs are given without knowing a proper target	3. Used to cluster the input data according to classes based on their statistical properties only.
4. Tools: • Neural Network • Support Vector Machine • Decision Tree	4. Tools: • Agglomerative hierarchical clustering • Divisive hierarchical clustering • K-means clustering • DBSCAN

2.4 Previous Research on Human Walking Path Analysis

2.4.1 An Approach for Extraction for Extraction of Human Walking Path in Intelligent Space – Hiromu Kobayashi, Hideki Hashimoto and Mihoko Niitsuma [10].

This journal explains on the method to extract human walking path by using Distributed Intelligent Network Device (DIND) and method to classify human walking path data. Distributed Intelligent Network Device is a fundamental of Intelligence Space (iSpace) which consists of sensors, processors and network devices. iSpace can perceive and understand events in whole space by communicating with each DIND. Figure 2.7 shows the concept of Intelligent Space.

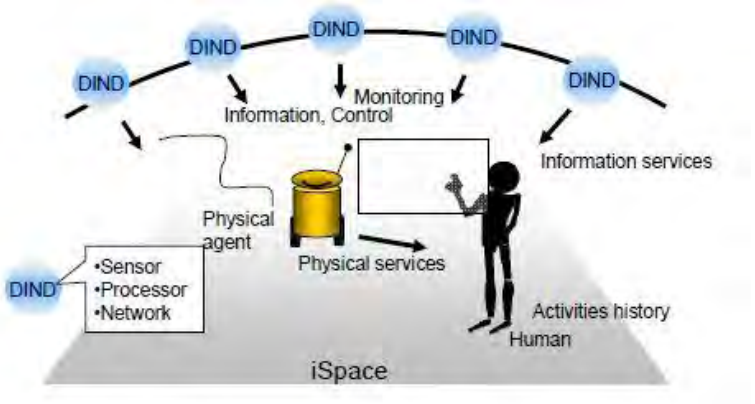


Figure 2.7: Intelligent Space concept [10].

During experiment, the sensor used is ultrasonic positioning system; which is divided into ultrasound receiver and ultrasound transmitter. During experiment, the subject will wears the transmitter which is mounted into student card strap, while the receiver is installed on the ceiling. The transmitter will regard the position of the subject by sending the signal to the receiver. 110 walking paths are obtained with consists of 5,481 points. Figure 2.8 shows the method on collecting human walking data and Figure 2.9 shows the path of subject resulting from the experiment.

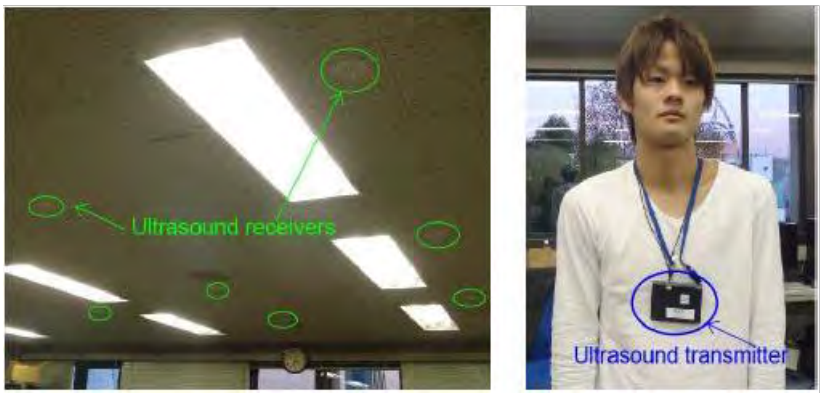


Figure 2.8: Ultrasound receivers which are installed to ceiling and ultrasound transmitter which is mounted to student card strap [10].

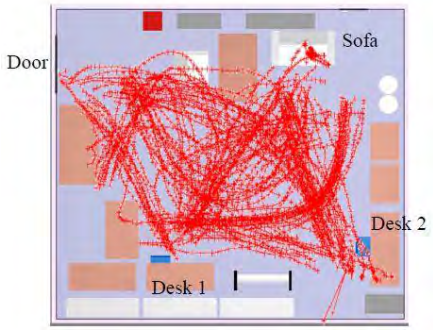


Figure 2.9: Experimental environment with obtained path [10].

As the human walking path data is not robust as is effected by human's stopped state, data smoothing method is used. In this paper, researchers used Generalized Cross-Validation (GCV) [11] based on empirical Bayes [12] approach in order to choose trade off parameter r. After that, the similarity is calculated to set an appropriate threshold to the dataset and to gain same information about the trajectories beforehand. In this process, Angular Metrics for Shape Similarity (AMSS) [13] is applied to compare similarities between two trajectories. During AMSS, Dynamic Programming is applied to maximize the total similarities.

The next is clustering the data. In this process, a hierarchical clustering method is used because the number of trajectories does not have to be specified.in hierarchical clustering, the medoid; trajectory with average minimum distance of same cluster is extracted and then clarify as representative trajectory in each cluster.