

SIGNATURE IDENTIFICATION BY HANDLING OUTLIERS WITHIN AUTHENTIC CASE BASE

Project Report Submitted in Partial Fulfilment of the
Requirements for the degree of
Master of Computer Application
Of
Jadavpur University
May, 2018

By

UTTAM KUMAR DAS

Master of Computer Application – III

Examination Roll Number: MCA186009

Registration Number: 133671 of 2015 – 2016

Under the guidance of

Dr. CHITRITA CHAUDHURI

Associate Professor

Department of Computer Science and Engineering

Faculty of Engineering and Technology

Jadavpur University

Kolkata – 700032, India

2018

**COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY**

TO WHOM IT MAY CONCERN

I hereby forward the project report entitled “*Signature Identification By Handling Outliers Within Authentic Case Base*” prepared by **Uttam Kumar Das** under my supervision to be accepted in partial fulfilment for the degree of **Master of Computer Application** in the Faculty and Technology of Jadavpur University, Kolkata.

(Dr. Chitrita Chaudhuri)
Associate Professor
Project Supervisor
Dept. of Computer Science and Engineering
Jadavpur University
Kolkata – 700032

Countersigned:

Prof. Ujjwal Maulik
Head, Dept. of Computer Science and Engineering
Jadavpur University
Kolkata – 700032

Prof. Chiranjib Bhattacharjee
Dean, Faculty of Engineering and Technology
Jadavpur University
Kolkata – 70032

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF APPROVAL *

The foregoing project report is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to the degree for which it has been submitted. It is understood that, by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project report only for the purpose for which it has been submitted.

Final Examination for
evaluation of the project

(Signatures of Examiners)

* Only in case the project report is approved

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this project report contains literature survey and original research work by me, the undersigned candidate, as part of my Master of Computer Application studies.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

NAME : Uttam Kumar Das

Examination Roll Number : MCA186009

Registration Number : 133671 of 2015 - 2016

Project Title : Signature Identification By Handling
Outliers Within Authentic Case Base

Signature with Date :

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of this task would be incomplete without the mention of the people who made it possible. Their constant guidance and encouragement crowned my effort with success.

It is a great pleasure to express my sincerest thanks to my project supervisor Dr. Chitrita Chaudhuri, Associate Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, for her encouragement, valuable suggestion, and constant support during the course of this project.

I would like to thank all the professors of the Department of Computer Science and Engineering, Jadavpur University, Kolkata for the guidance they provided me throughout the duration of the Master of Computer Application course.

A special note of thanks goes to Prof. Ujjwal Maulik, Head, Department of Computer Science and Engineering, Jadavpur University.

I am also thankful to Prof. Chiranjib Bhattacharjee, Dean, Faculty of Engineering and Technology, for providing an excellent environment for completion of this project.

I am also indebted to my co-researchers Ms. Shisna Sanyal, Mr. Anupam Baidya and Mr. Saunak Roy Chowdhury for their seamless co-operation and help in completion of this project. I am thankful to my fellow classmates and my family for constant help and support.

Date: _____

Uttam Kumar Das
Master of Computer Application – III
Examination Roll No. – MCA186009
Registration No: 133671 of 2015 - 2016

CONTENTS

1	Introduction	1 - 2
2	Basic Concepts	3 - 8
2.1	Machine Learning	3
2.2	Classification	3
2.3	Case Based Reasoner (CBR)	4 - 5
2.3.1	Brief Discussion	4 - 5
2.3.2	Present perspective	5
2.4	Digital Image Processing	6
2.5	Distance Measures	7
2.6	Statistical Measures	8
2.7	Outlier	8
3	Methodology	9 - 17
3.1	Description of Input	9 - 11
3.1.1	Attribute sets	9 - 11
3.1.2	Discretization Techniques	11
3.1.3	Partitioning of Training and Test datasets	11
3.2	Description of storage structure	12
3.3	Description of adopted Algorithm	13 - 17
3.3.1	Algorithm: Identification Accuracy	13 - 17
4	Experimental Setup	18 - 19
4.1	Datasets	18 - 19
4.2	Machine Configuration	19

5	Results and Performance Analysis	20 - 23
5.1	Overall Performance	20 - 21
5.2	Bar Charts of Accuracy Percentage	21 - 23
5.2.1	Dataset – 1	21
5.2.2	Dataset – 2	22
5.2.3	Dataset – 3	23
6	Conclusion and Future Scope	24
	Bibliography	25

List of Tables

Table No.

1	Accuracy Percentage for different datasets	20
---	--	----

List of Figures

Figure No.

1	CBR life cycle	5
2	Structure of the case-base	5
3	Digital image processing : Layout	6
4	A right angled triangle	7
5	Authentic Signature List details	12
6	Structure of the feature-base	12
7	Sample signature of Dataset 1	18
8	Sample signature of Dataset 2	18
9	Sample signature of Dataset 3	19
10	Bar chart for Dataset – 1	21
11	Bar chart for Dataset – 2	22
12	Bar chart for Dataset – 3	23

Chapter 1

Introduction

The handwritten signature of any person is an important biometric characteristic which is used prevalently for identification in financial and business transactions. Invigilation in examination halls are also conducted via signature identification and authentication. System fraudulence and unauthorised intrusion are crimes which are on the rise these days. One time-tested method of protection is by the use of an identification system based on handwritten signatures. This method is cost-efficient and simple as compared to other bio-metric methods.

Handwritten signature, strictly defined, is full or part of a name written in ones own handwriting. But in reality signatures are composed of special characters and flourishes and therefore most of the time they can be unreadable.

Signature identification systems are broadly classified in two ways: on-line and off-line. As the name signifies, on-line signatures are directly signed onto the electronic tablets or the screen using some appliances like the digital pen. Thus the technique involves sophisticated and costly tools.

Here we are restricting ourselves to off-line techniques only for person identification. In this technique, a signature is made on an external media such as a piece of paper and scanned to produce a digitized copy to be stored within a computer system. Due to the relative ease of use of an offline system, a number of applications worldwide prefer to use this system for person verification (e.g. Cheque verification in any bank).

In the proposed system, the preserved signatures of an individual are used to train the machine to recognize the person when an authentic signature of the same person is presented. The stress is on the word authentic, as fraudulence does not play a part in person verification. Signatures are preserved as attributes or features extracted from the individual's signature images. These features are considered as the input, which are then processed based on some predefined standardized methods to produce a class value. The class value gives the identity of the signatory.

So the present work is oriented towards building a classifier that helps to detect the identity of a person based on the person's off-line signatures which are already preserved in a base. Ideally, the classifier needs to maintain the preserved signatures in a manner which enhances the knowledge base of the system, keeping it in a constantly updated state. We have studied and employed some of the techniques associated with Case-Based Reasoning (CBR) for this purpose.

The features themselves must be in proper order as they are the determining factor on which recognition depends. Hence the need for outlier detection amongst the features and handling the same using some prevailing statistical techniques.

In chapter 2, we introduce some concept of Machine Learning, Classification, CBR, Image processing and definition of other parameters deemed important in the present work such as distance measures, statistical measures and outliers. In chapter 3 is described the methodologies used to determine Identification Accuracy based on different attribute sets. In chapter 4, we discuss about the datasets used, as well as the system configuration and tools utilised for our work. Chapter 5 describes the results obtained from the experimental procedures carried out during the research and provide summaries on the basis of datasets and attribute sets used. Lastly chapter 6 provides the conclusion drawn on the outcome of the experiments. This chapter also hints at future scopes in this research domain.

Chapter 2

Basic Concepts

2.1 Machine Learning

Machine learning, which is an application of artificial intelligence (AI), provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it to enhance its knowledge base. The process of learning begins with data in order to look for patterns and make better decisions in the future, based on the data that we provide. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly [1].

Machine learning algorithms can be broadly classified into two categories, namely Supervised learning and Unsupervised learning. Supervised learning is a learning in which we train the machine using data (designated as training data) which is well labelled. After that, the trained system is provided with a new set of data (designated as test data) in order to evaluate the label of the data correctly. Unsupervised learning is the process of building a system using data that is neither classified nor labelled and allowing the algorithm to act on that data without guidance. Here the job of the machine is to group the data according to similarities, patterns and differences without any prior knowledge of class value.

Some areas where machine learning is used are biometric identification, computer vision, game playing, Natural language Processing (NLP), recommendation system, financial market analysis to name a few.

In the present work, a classifier is utilised to identify a person by comparing distances between vectors representing offline handwritten signature images of that person.

2.2 Classification

Classification is the process of supervised learning which identifies the category of a new instance presented to the classifier [2]. The technique demands that the data with known class value be divided into two parts – the training set and the test set. The classifier model is built using the training dataset. Now this model is tested to obtain a class value for each of the test data tuples, which when checked against the known class value of that tuple helps to determine the accuracy attained by the classifier. Major classification techniques include Decision Trees, Bayesian Classifiers, Rule Based Classifiers, Neural Networks, Support Vector Machines, K – nearest Neighbour Algorithm, and Case-Based Reasoners (CBR).

This last classification technique, CBR, mentioned above have been employed over here to identify a person using authentic signature bases.

2.3 Case-Based Reasoner (CBR)

2.3.1 Brief Discussion:

CBR classifiers treat every problem-solution pair as a case and each such case is stored in a base. An unsolved problem is supplemented with its correct solution which represents its class value. Often, a case base, besides a detailed statement of the problem and its solution, also houses the necessary meta-data required for the problem.

As mentioned in the work by U. Farhan et.al. [3], CBR brings some important advantages to the problem-solving strategy. It can reduce the processing time significantly and also be very useful when domain knowledge is not completely available or not easy to obtain, although extensive knowledge and expertise in the field always helps while modifying the similar solutions to produce a new solution. It also accommodates incremental learning techniques by allowing new cases to upgrade the system knowledge overall.

Most importantly, potential errors can be avoided and past mistakes rectified in similar cases, while attending to problem at hand. Search time may be reduced by a fool-proof indexing technique.

Thus CBR is an artificial intelligence (AI) technique that considers old cases to take decision for new situations. The old cases constitute past experiences on which one can rely, rather than on rules, during the decision making process. CBR works by recalling similar cases to find solution to new problems [3].

To insert a problem-solution pair in the case-base, at first we search for that particular problem within the existing base with the help of some predefined indexing system. If an exact match is found for the present problem, there is no need for insertion. Otherwise, cases constituting the nearest matches are found, and their class information are collected to provide the new case with a suitable class value. The new case with its new solution is now ready for insertion into the case-base.

The CBR processes include four main steps [4]:

Retrieve

Given a target problem, *retrieve* from memory cases relevant to solving it. For fast retrieval the pre-requisite is an efficient indexing technique.

Reuse

The solution(s) from the retrieved case(s) need to be mapped to the target problem. This may involve adaptation or *reuse* of the solution(s) as needed to fit the new situation.

Revise

Having mapped the retrieved solution(s) to the target situation, the new solution generated has to be tested in the real world (or a simulation) and, if necessary, need to be *revised*.

Retain

After the solution has been successfully adapted and revised for the target problem, *retain* the resulting experience as a new case in memory.

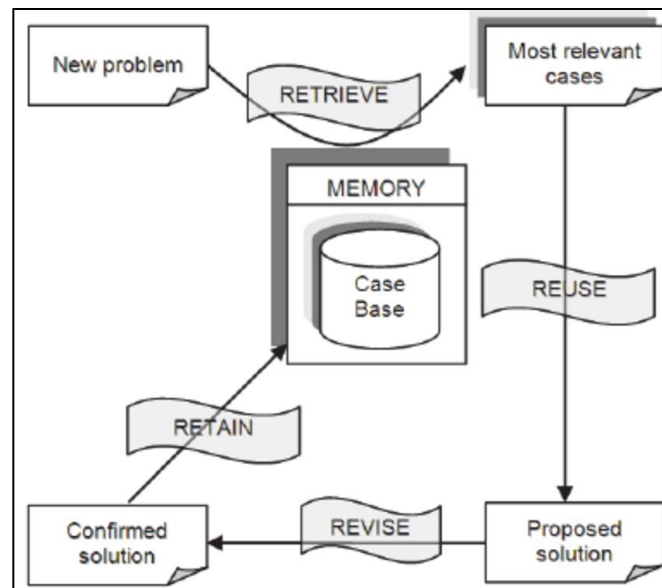


Figure 1 CBR life cycle [5]

2.3.2 Present perspective:

Here CBR is being utilized to establish a person's identity with the help of offline digitized signatures, the most prevalent mode of biometric authentication. A fresh signature of the person is taken and matched against the authentic signatures already preserved within the system in the form of a case-base and the reasoner seeks to identify the person as accurately as possible. The process demands that at the time of registration (for example while creating an account in a bank or applying for inclusion in a societal group), a set of authentic signatures be taken as proof of the applicant's identity. These signatures will be scanned, digitized and pre-processed prior to extraction of feature values in the form of discrete numbers. The discretized feature values can now be stored in a case-base with other general relevant information about the person, to introduce a starting expertise level within the knowledge-base.

Person 1	Authentic Signature list	General information
Person 2	Authentic Signature list	General information
:	:	:
Person n	Authentic Signature list	General information

Figure 2 Structure of the case-base

Detailed description of the case-base structure is given in chapter 3.

2.4 Digital Image Processing

The field of digital image processing refers to processing digital images by means of a digital computer. A digital image is composed of a finite number of elements, each of which has a particular location (provided by the dimensions of the point) and value (the intensity level at that point). These elements are called picture elements or pixels [6].

In the present context, image processing involves low-level processes and mid-level processes. Low-level processes include primitive operations such as image pre-processing to reduce noise, contrast enhancement, and image sharpening. Mid-level processing on images involves tasks such as segmentation (partitioning an image into regions or objects), description of those objects to reduce them to a form suitable for computer processing, and classification (recognition) of individual objects. A mid-level process is characterized by the fact that its inputs generally are images, but its outputs are attributes extracted from those images [6]. The following Figure 3 depicts the different steps associated with a typical digital image processing task.

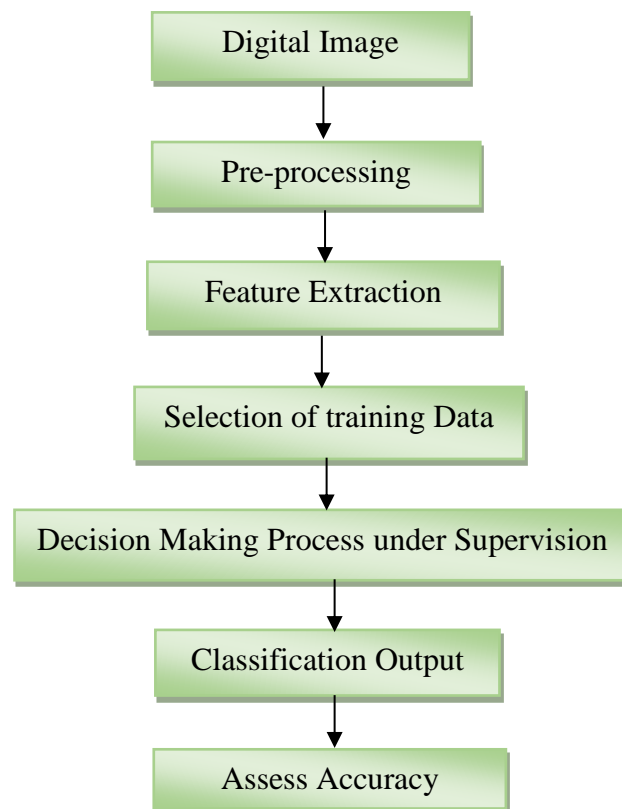


Figure 3 Digital image processing : Layout

Use of image processing techniques is increasing day by day in all spheres of life. In our case, we deal with an offline handwritten signature image, scanned, digitized and pre-processed (Figure 7, 8 and 9 in section 4.1) to produce datasets of numeric features. Then we have partitioned these datasets into training and test datasets, as described in section 3.1. The classifier model is next built to identify a person's signature, as discussed in section 3.3 and its accuracy assessed based on results depicted in chapter 5.

2.5 Distance Measures:

The distance measures give us the proximities between objects [7]. In our identification problem, we treat each signatures as a vector in an n – dimensional space, where every element of the vector is represented by one of the n feature values. So to calculate the similarity (or proximity) between two signature images i.e., two points in n – dimensional space, we utilize different distance measures.

Minkowski distance [7] between two points $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$ is defined as:

$$d = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} \dots\dots\dots (1)$$

, where p is a positive integer, n is the number of elements in the feature vector, and x_i and y_i are the values of the i 'th element of X and Y respectively.

Euclidian distance and Manhattan distance are specialized forms of Minkowski distance. When $p = 2$, it gives the Euclidean distance and for $p = 1$, Manhattan distance is obtained. From past researches in this domain, the Manhattan distance has been found to give the best result. So, in this work we have used Manhattan technique of finding the distance between two signatures using the following formula derived from equation (1) above:

$$d = \sum_{i=1}^n |x_i - y_i| \dots\dots\dots (2)$$

These distance measures follow the famous Pythagoras theorem from Euclidean geometry as depicted in the following Figure 4 of a right angle triangle.

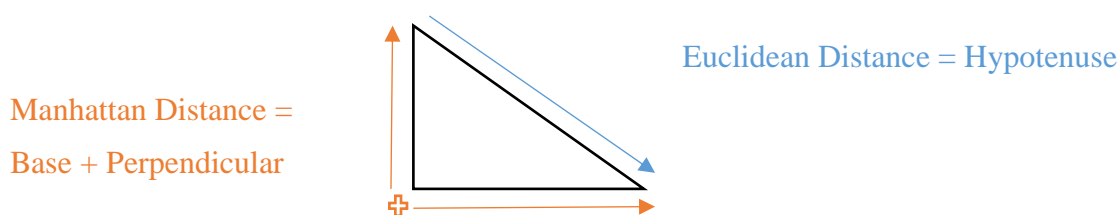


Figure 4 A right angled triangle

2.6 Statistical Measures:

There are three prevalent central tendency measures for all numeric data – the **mean**, the **median** and the **mode**. The **mean** is the average of the dataset obtained by summing up all the values and dividing the same by the cardinality of the set. The **mode** is the most the commonly recurring value within the dataset.

The **median** is the middle number in a group of numbers (data points). It's not as commonly used as other central tendencies such as **mean** and **mode**. But in many practical situations, it can be the best 'average' to use when you have a set of data that contains outliers [8].

Steps for finding the median in a group of data points are as follows:

- The data points are first arranged in ascending order of magnitude.
- Total number of data points is counted.
- There are two possibilities depending on the total number of data points.
- If the total count is odd, then the median will be the one exactly in the middle, with an equal amount of points on either side of it.
- If the total count is even, then there will not be a single point in the middle. In this case, the average of the two middle points is taken as the median value.

Data is often represented by a five numbers summary – the most commonly used percentile values given by the first quartile ($Q1$) / 25th percentile; the second quartile ($Q2$) / the median / 50th percentile; the third quartile ($Q3$) / 75th percentile and the minimum and the maximum. The method used to generate the minimum and the maximum values is dependent on a simple measure of spread, provided by a distance called the Inter Quartile Range (IQR), defined as $IQR = Q3 - Q1$. A common rule of thumb for identifying suspected outliers is to single out values falling at least $1.5 \times IQR$ above the third quartile or below the first quartile. So these two points are marked as the lower and upper limits of the tolerable range [7].

2.7 Outlier:

There exist data objects that do not comply with the general behaviour or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called **outliers** [7]. They are statistical observations that are numerically distant from the rest of the data. They can skew the mean of the data, but by using the median, statisticians can get a more accurate picture of the true average (or middle value) of the data [8].

In our work, we have picked out the outlier attribute values from the raw dataset, based on the statistical technique involving Inter Quartile Range (IQR) value obtained for each individual attribute, as discussed in section 3.3 below. Then in some of the techniques described in the same section, we have capped these outliers to the nearest boundary of the tolerable range.

Chapter 3

Methodology

The success of a project depend mainly on the outcome obtained from experimental results. The outcome, in its turn, is dependent on proper input and correct techniques employed on those input – basically the concept of Garbage-In Garbage-Out (GIGO).

In this chapter, we describe the input used in our experiments as well as the technologies utilized to provide the appropriate output. We start with a general definition of attribute followed by a detailed description of the data attributes availed and the algorithms adopted.

Definition of attribute:

An attribute is an aspect of an instance or a specification that defines a property of an object. Attributes are often called features in Machine Learning. A special attribute is the class label that defines the class the instance belongs to [9].

3.1 Description of Input:

For representing an offline handwritten signature image, we have used different types of attributes belonging to two broad classes – Global attributes and Local attributes.

An attribute that comes from the signature as a whole, is termed as a global attribute.

An attribute that is constructed from one part of the signature image, is termed as local attribute.

In our experiments, we have utilized one set of global attributes and one set of local attributes, individually and in combination. A description of each of these sets is given in the following sub-section.

3.1.1 Attribute sets:

Here three sets of attributes are being considered – **A**, **B** and **C**. **A** represents the global attribute set and **B** represents the local attribute set. **C** is a combination of both **A** and **B**. All these attributes have been well tested during prior research in the same domain carried out in the university laboratories. A list of these attribute sets is given below:

A. Attribute set 1 (Global Attributes) [13] [14] [15]:

1. Signature height
2. Pure width
3. Image area
4. Vertical centre
5. Horizontal centre
6. Maximum vertical projection
7. Maximum horizontal projection
8. Number of vertical projection peaks
9. Number of horizontal projection peaks
10. Baseline shift
11. Number of edge points
12. Number of cross points
13. Number of closed loops
14. Top heaviness
15. Horizontal dispersion
16. Interior to outline pixel ratio
17. Mean ascending height
18. Mean descending height
19. Reduced no of components
20. Number of significant components
21. Global slant
22. Local slant
23. Mean slant

B. Attribute set 2 (Local attributes):

It contains 40 angular distance features, indigenously developed by a prior researcher of this university, constructed as follows:

- I. At first, the binarized, cropped and thinned signature image is re-sized to a dimension of 400 x 100 pixels.
- II. Then the signature image is divided into 4 vertical parts, each of 100 x 100 dimension. The following steps are carried out for each of these vertical parts:
 - a. The centre of mass is calculated for the part.
 - b. Next we divide the part into 8 equal angular regions by considering radial lines from the centre of mass outwards, each at an angular interval of 45° .
 - c. Now we find the farthest black pixel within each of these 8 angular regions, and compute the Euclidean distance of these pixels from the centre of mass. If

there is no black pixel present in any one of these regions, then the corresponding distance value is set to 0. So, we have 8 distance values per part.

- d. Next we calculate the geometric centre of the part, and we compute the Euclidean distance of this geometric centre from the centre of mass of that part. This constitutes the ninth attribute value per part.
- e. Lastly, the tenth attribute of the part is obtained by calculating the angle with the horizon subtended by the line joining the centre of the mass and the geometric centre of the part.

III. Thus we have 10 local attributes for each of the four parts – i.e., 40 attributes in all for the total signature image.

C. Attribute set 3 (Global features + Local features):

We merge the attributes in attribute set 1 and attribute set 2, to generate the combined set of attributes. Thus, this set contains a total of 63 attributes, consisting of 23 global attributes and 40 angular distance attributes.

3.1.2 Discretization Technique:

Discretization, as a pre-processing step for data mining, is a process of converting the continuous attributes of a data set into discrete ones so that they can be treated as nominal features by a machine learning algorithm [10].

Here the raw data of the signatures are discretized, so that if the discretization level be x for a dataset, then all its feature values lie in the range 0 to $x - 1$. The actual techniques used are described in the next sub-section 3.3 below.

3.1.3 Partitioning of Training and Test Datasets:

For determining identification accuracy, only authentic signatures of persons are required, from which a major portion is taken as the training set and the rest is utilized to test the model built. The actual ratio between the training and the test set generally depends on data size. An ideal combination is a 2:1 ratio between training and test dataset.

3.2 Description of storage structure:

For easy access and updation, we are preserving the detail of each person as a case in the case-base, the outline structure of which has already been shown in Figure 2 within chapter 2. The detail composes of an index value, authentic signatures list and other general information about each person. As depicted in the following Figure 5, for every signature, there is a field for storing all the discretized feature values for that signature and the time stamp information (if available). Based on the feature values thus obtained, for all the signatures, and for all the persons, we next generate statistical information for each feature viz. Q1, Q2, Q3, IQR and tolerable range (as already discussed), which are stored separately in a feature-base shown in Figure 6 below.

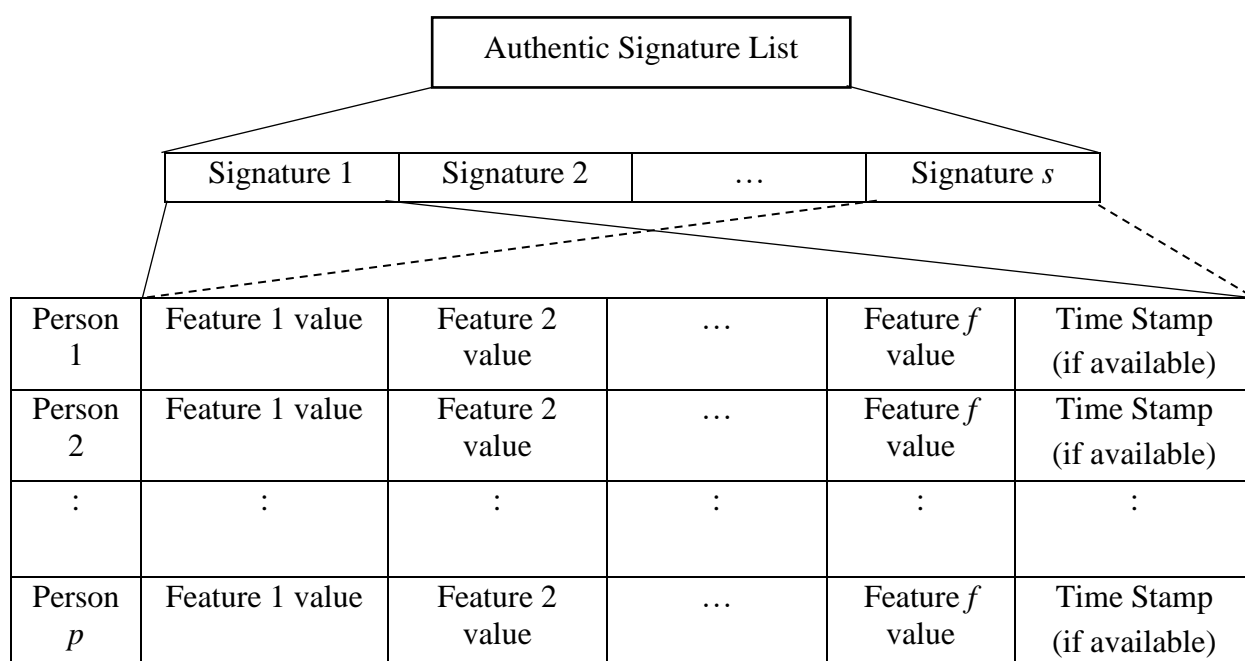


Figure 5 Authentic Signature List details

Feature 1	Q1	Q2	Q3	IQR	Tolerable Range
Feature 2	Q1	Q2	Q3	IQR	Tolerable Range
:	:	:	:	:	:
Feature f	Q1	Q2	Q3	IQR	Tolerable Range

Figure 6 Structure of the feature-base

3.3 Description of adopted Algorithm:

As we know, no two signatures of a person will be exactly the same. So, we depend on the nearest signature image matching for identifying the person. The outline of the strategy is next discussed briefly.

There are two types of discretization technique adopted here. The first involves finding the lower and upper limits of the tolerable range for each attribute of the training dataset depending on the IQR obtained for the attribute. The training and test datasets are then discretized using this tolerable range. The second technique sets the range between the physical minimum and maximum of the training dataset values for each attribute, and during discretization this range is utilized in both training and test datasets.

We have adopted four methods for employing capping of data during experimentation, i.e., setting out of bound attribute values to the lower and upper limits of the tolerable range. In the first three of these methods, described as Method – 1, Method – 2 and Method – 3, we have utilized the first discretization technique based on IQR calculation.

In Method – 1, capping is applied only in the training set, followed by discretization of training and test datasets thereafter. Method – 2 is the same as Method – 1 except that the capping is applied on both the training and test datasets here. In Method – 3, no capping is applied either on the training or the test dataset, other processes remaining the same as Method 1 and 2.

Method – 4, where the second discretization technique is adopted, the range value is used to discretize both training and test datasets without any capping.

Now we employ classification techniques based on Manhattan distance calculated between each test data and the whole of training dataset to predict the identity of the test person. The details of the technique is given in the algorithm described in the following sub-section. The accuracy of the classifier is measured in terms of matched identification within the test dataset.

3.3.1 Algorithm: Identification Accuracy

Input:

- P – Total number of persons in the dataset
- F – Total number of features
- S – Total number of authentic training signatures per person
- ST – Total number of test signatures per person
- $FV[P, S, F]$ – The corresponding feature vector
- $MAX[F]$ – Vector storing maximum value for all features
- $MIN[F]$ – Vector storing minimum value for all features
- D – Discretization value
- $LARGE = 9999$

Output:

- Percentage of Identification accuracy

Main_Method

BEGIN

1. Call *Method – 1*
2. Call *Method – 2*
3. Call *Method – 3*
4. Call *Method – 4*

END

Method – 1

BEGIN

1. Call *Max_Min_IQR ()*
2. Call *Feature_Capping ()*
3. Call *Discretization ()*
4. Call *Identification_Accuracy ()*

END

Method – 2

BEGIN

1. Call *Max_Min_IQR ()*
2. Call *Mod_Feature_Capping ()*
3. Call *Discretization ()*
4. Call *Identification_Accuracy ()*

END

Method – 3

BEGIN

1. Call *Max_Min_IQR ()*
2. Call *Discretization ()*
3. Call *Identification_Accuracy ()*

END

Method – 4

BEGIN

1. Call *Max_Min ()*
2. Call *Discretization ()*
3. Call *Identification_Accuracy ()*

END

Procedure *Max_Min_IQR* ()

BEGIN

1. For each $f = 1$ to F
 2. Set $t = -1$
 3. For each $p = 1$ to P
 4. For each $s = 1$ to S
 5. $t = t + 1$
 6. $M[t] = FV[p, s, f]$
 7. EndFor
 8. EndFor
 9. Vector M is sorted in ascending order.
 10. Set $Q1[f]$ to the first Quartile or 25th percentile value of M
 11. Set $Q2[f]$ to the 2nd Quartile or 50th percentile value of M
 12. Set $Q3[f]$ to the 3rd Quartile or 75th percentile value of M
 13. Set Inter-Quartile Range $IQR[f] = Q3[f] - Q1[f]$
 14. Set Lower-Limit $MIN[f] = Q1 - 1.5 * IQR[f]$
 15. Set Upper-Limit $MAX[f] = Q3 + 1.5 * IQR[f]$
 16. EndFor
- END

Procedure *Max_Min* ()

BEGIN

1. For each $f = 1$ to F
 2. Set $t = -1$
 3. For each $p = 1$ to P
 4. For each $s = 1$ to S
 5. $t = t + 1$
 6. $M[t] = FV[p, s, f]$
 7. EndFor
 8. EndFor
 14. $MIN[f] =$ Minimum Value within M
 15. $MAX[f] =$ Maximum Value within M
 16. EndFor
- END

Procedure *Feature_Capping* ()

BEGIN

1. For each $p = 1$ to P
 2. For each $s = 1$ to S
 3. For each $f = 1$ to F
 4. If $FV[p, s, f] < MIN[f]$ then
 5. $FV[p, s, f] = MIN[f]$
 6. EndIF
 7. If $FV[p, s, f] > MAX[f]$ then
 8. $FV[p, s, f] = MAX[f]$
 9. EndIF
 10. EndFor
 11. EndFor
 12. EndFor
- END

Procedure *Mod_Feature_Capping* ()

BEGIN

1. For each $p = 1$ to P
 2. For each $s = 1$ to $S + ST$
 3. For each $f = 1$ to F
 4. If $FV[p, s, f] < MIN[f]$ then
 5. $FV[p, s, f] = MIN[f]$
 6. EndIF
 7. If $FV[p, s, f] > MAX[f]$ then
 8. $FV[p, s, f] = MAX[f]$
 9. EndIF
 10. EndFor
 11. EndFor
 12. EndFor
- END

Procedure *Discretization* ()

BEGIN

1. For each $p = 1$ to P
 2. For each $s = 1$ to $S + ST$
 3. For each $f = 1$ to F
 4. If $MAX[f] == MIN[f]$ then
 5. $FV[p, s, f] = 0$
 6. Else
 7. $FV[p, s, f] = int\left(\frac{(FV[p, s, f] - MIN[f])}{MAX[f] - MIN[f]} * D\right)$
 8. EndIF
 9. EndFor
 10. EndFor
 11. EndFor
- END

Procedure *Identification_Accuracy* ()

BEGIN

1. Set $count = 0$
 2. For each $p = 1$ to P
 3. For each $s = 1$ to ST
 4. $U \equiv$ Test signature s of person p
 5. Set $dist = LARGE$
 6. For each $p1 = 1$ to P
 7. For each $s1 = 1$ to S
 8. $V \equiv$ Authentic signature $s1$ of person $p1$
 9. $d =$ Manhattan distance between signatures U and V
 10. If $d < dist$ then
 11. $dist = d$
 12. $id = p1$
 13. EndIF
 14. EndFor
 15. EndFor
 16. If $id == p$ then
 17. $count = count + 1$
 18. EndIF
 19. EndFor
 20. EndFor
 21. Evaluate $accuracy = \left(\frac{count}{P * ST}\right) * 100$
- END

Chapter 4

Experimental Setup

4.1 Datasets:

Here we have used 3 sets of signature data for proving the viability of the system. Following is a descriptions of the datasets.

1. Dataset 1 (Designated as OUR Dataset):

As the name suggests, this signature set was collected indigenously from 121 volunteers, mainly university students and their acquaintances. Each person provided us with 20 authentic signatures. Participants in this program were asked to sign using black ball-point pen having 0.5 mm tip. The space demarcated for each signature was a rectangular box of 9 cm x 3 cm area. All the signatures in this set were scanned at a resolution of 200 dpi to obtain grey scale images.

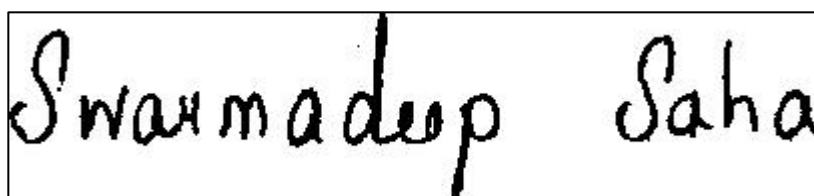
A rectangular box containing a handwritten signature in black ink. The signature reads "Swarnadep Saha" in a cursive, slightly slanted script.

Figure 7 Sample signature of Dataset 1

2. Dataset 2 (Designated as ATVS Dataset):

This one consists of 2250 signatures belonging to the standard MCYT Bimodal Biometric Database [11] scanned at a resolution of 300 dpi with 15 genuine and 15 skilled forgeries for each of 75 persons. The signatures were obtained as grey-scale images.

A rectangular box containing a highly stylized, abstract handwritten signature in black ink. The signature is composed of several overlapping loops and sharp peaks, making it difficult to decipher.

Figure 8 Sample signature of Dataset 2

3. Dataset 3 (Designated as Persian Dataset):

It is a Persian Dataset, obtained from the website [12]. This consists of the off-line handwritten signatures of 115 persons, with 27 genuine signatures per person. Overall 3105 authentic images were collected from graduate students of University of Tehran and Sharif University of Technology. Signatures were scanned with 600 dpi resolution. This dataset was included in the test to establish the role of image processing in our research work. The accuracy obtained with this signature set proves that each signature is treated by our system as a pattern irrespective of the language of the signatory.

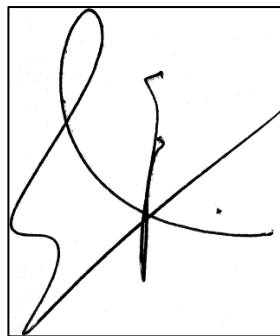


Figure 9 Sample signature of Dataset 3

Here we worked with only the genuine signatures of each dataset.

4.2 Machine Configuration:

System – Lenovo™ ideapad 320

Processor: Intel® Core™ i3-6006U CPU @ 2.00GHz
RAM: 4.00 GB
System type: Windows 10 Enterprise, 64-bit Operating System, x64-based processor

Tools – MATLAB R2017a

Chapter 5

Results and Performance Analysis

5.1 Overall Performance:

Table – 1: Accuracy Percentage for different datasets

Dataset Name	No. Of Persons	No. Of Authentic Signatures	No. Of Test Signatures	Attribute Set	Method	Discretization Value (D)	Percentage Of Accuracy
Dataset – 1 (OUR)	121	15	5	1 (Only Global)	Method – 1	27	83.96694215
					Method – 2	27	83.96694215
					Method – 3	27	82.97520661
					Method – 4	33	82.6446281
				2 (Only Local)	Method – 1	19	91.90082645
					Method – 2	19	91.90082645
					Method – 3	25	91.73553719
					Method – 4	28	93.55371901
				3 (Combined)	Method – 1	12	97.19008264
					Method – 2	12	97.19008264
					Method – 3	13	97.19008264
					Method – 4	15	97.19008264
Dataset – 2 (ATVS)	75	10	5	1 (Only Global)	Method – 1	22	67.2
					Method – 2	22	67.2
					Method – 3	25	64.8
					Method – 4	33	69.6
				2 (Only Local)	Method – 1	30	69.33333333
					Method – 2	30	69.33333333
					Method – 3	30	67.2
					Method – 4	25	70.13333333
				3 (Combined)	Method – 1	22	82.66666667
					Method – 2	22	82.66666667
					Method – 3	29	80
					Method – 4	31	81.33333333
Dataset – 3 (PERSIAN)	115	20	7	1 (Only Global)	Method – 1	23	70.0621118
					Method – 2	23	70.0621118
					Method – 3	23	66.83229814
					Method – 4	33	65.59006211
				2 (Only Local)	Method – 1	20	74.65838509
					Method – 2	20	74.65838509
					Method – 3	28	74.9068323
					Method – 4	17	76.39751553
				3 (Combined)	Method – 1	17	87.57763975
					Method – 2	17	87.57763975
					Method – 3	23	84.8447205
					Method – 4	22	83.47826087

For all the datasets, the combined Attribute set – 3 gives the best accuracy, as apparent from the highlighted figures in the above table 1.

For each dataset the performance evaluations are graphically represented by bar charts displayed below in section 5.2.

5.2 Bar Charts of Accuracy Percentage:

5.2.1 Dataset – 1

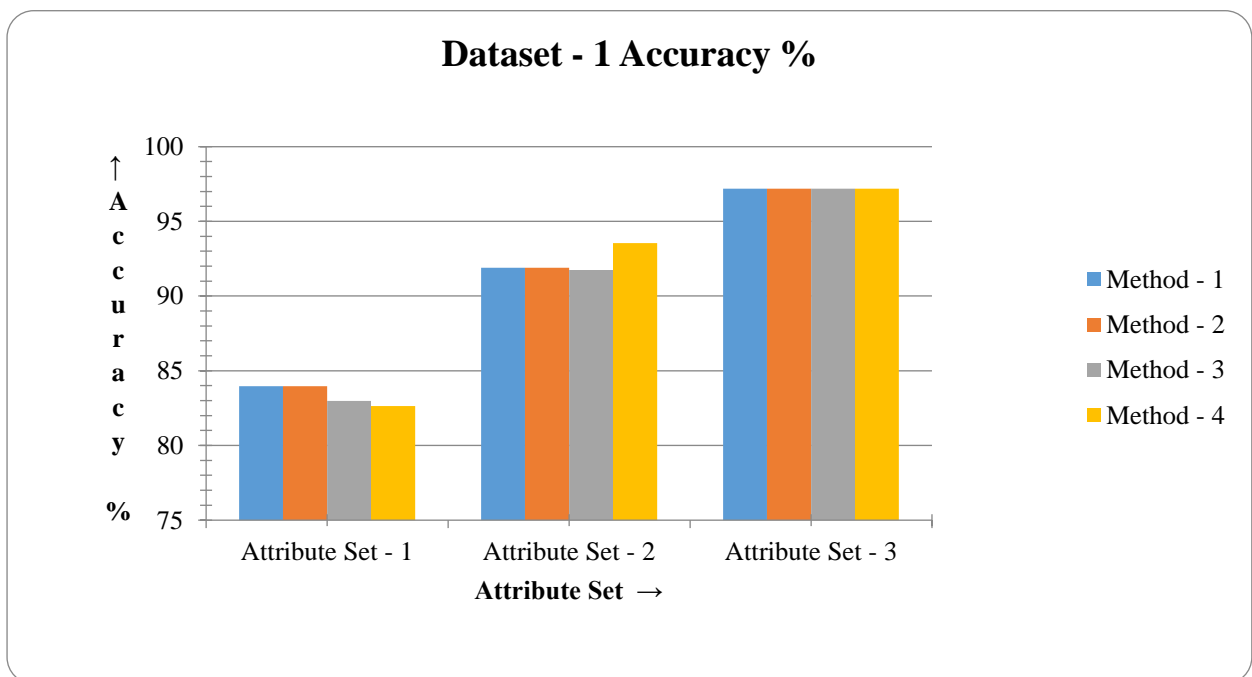


Figure 10 Bar chart for Dataset – 1

For dataset – 1, considering only the global attributes (attribute set – 1), gives very poor results as is seen in the above bar chart (Figure – 10). The local attributes (attribute set – 2), comprising of the angular distance features, improves the accuracy level to an extent. Among all these three, attribute set – 3, which contains all global and local attributes, gives the best accuracy. Surprisingly, all four methods are found to be equally effective.

5.2.2 Dataset – 2

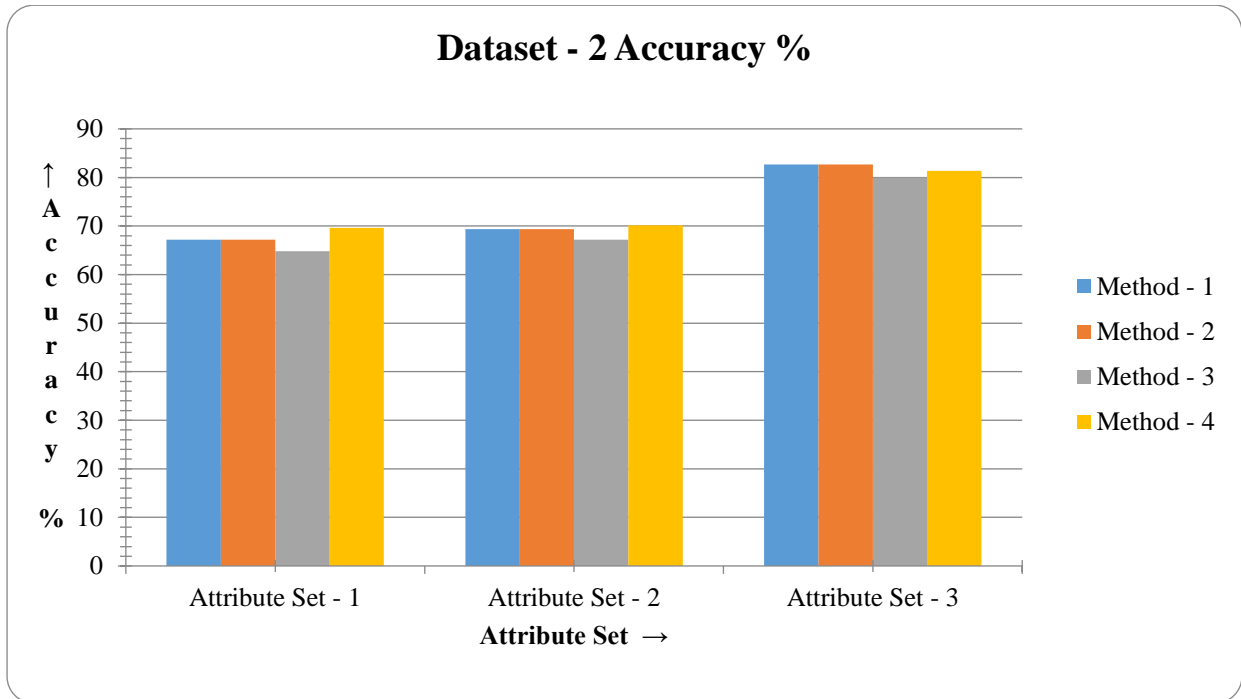


Figure 11 Bar chart for Dataset – 2

For Dataset – 2, the performance remains almost the same in case of using global and local attributes individually. Here the best accuracy levels are obtained with Method – 4, which utilises minimum and maximum values for tolerable range calculation in the training set to normalize both training and test datasets, without any capping. Method – 1 and Method – 2, both of which involve spread of data calculated over the training set IQR, give the next best results for the individual global and local attributes. The two methods differ in the capping associated, which is effective on only training for Method – 1 and on both training and testing for Method – 2. Method – 3, which is also based on IQR, but does not apply capping on either training or test datasets, is the poorest performer of the lot.

The best accuracy overall is attained again for the combination attribute set – 3, where Methods 1 and 2 are the winners and Method – 4 is the runners up.

5.2.3 Dataset – 3

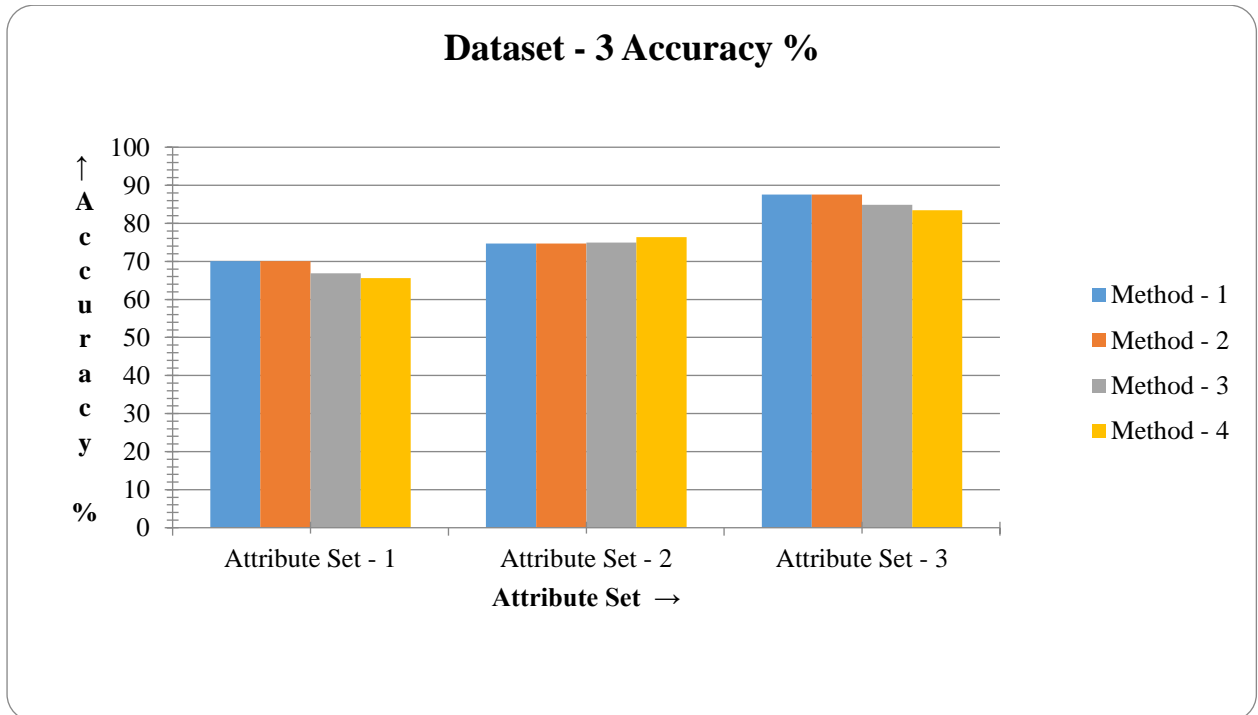


Figure 12 Bar chart for Dataset – 3

Global and local attributes perform differently for Dataset – 3, as is evident from the above bar chart (Figure 12). But the overall best trend of the combined attribute set – 3 remains the same and the performance of Methods 1 and 2 is still the highest as in both earlier cases. There is a slight discrepancy in the performance of Method – 4 with an anomalous winning streak observed for the local attributes alone.

Chapter 6

Conclusion and Future Scope

For all three datasets we have found the third (combined) attribute set to be most effective in increasing the identification accuracy level of the classifier using discretized training and test datasets, based on IQR calculation with capping of attribute values within the training set (Method 1). The effect of capping attribute values within the test dataset (Method 2) does not seem to affect the accuracy level achieved through Method 1 either way, although this may be a peculiarity of the datasets used.

The effect of not capping the outlier attributes is detrimental to the performance of the system, which is proved almost universally by the results shown in Method 3 for each of the datasets and for almost all combinations of attributes.

The usage of minimum and maximum values of attributes for normalization within training dataset and later applying the same for the test dataset, without involving any capping, as in Method 4, lowers down the accuracy level slightly for both datasets 1 and 2 and drastically for the third dataset.

So the usage of statistical techniques, calculated over the training set with capping applied on the same to effectively handle outlier attributes, is clearly indicated for best performance overall. As already stated, the Manhattan distance measure was selected for comparison based on the findings of earlier research work done on the same sets of data. For the indigenous dataset, we achieve 97.19 % of identification accuracy utilizing these techniques.

A future scope in this domain may involve updation of the case-base by inserting newly identified signatures either by appending (if there is scope of append) or in lieu of existing ones otherwise. The replacement of old signatures can be done in two ways. The first one would involve checking the difference in the time stamps of the two signatures and replacing the old one with the new only if the difference exceeds some predefined threshold time period. This was not possible to be tested with the existing datasets as Time Stamp information was not available with them. The second method would compare proximity of the median signature of a person with all the existing authentic signatures of that person, and replace the most distant one with the new, if it is nearer.

Bibliography

- [1] <http://www.expertsystem.com/machine-learning-definition/>
- [2] https://en.wikipedia.org/wiki/Statistical_classification
- [3] Uday Farhan, Majid Tolouei-Rad, and Adam Osseiran. Indexing and retrieval using case-based reasoning in special purpose machine designs. *The International Journal of Advanced Manufacturing Technology*, pages 1 – 15, 2017.
- [4] Ian Watson. *Applying Case-based Reasoning: Techniques for Enterprise systems*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.
- [5] https://www.researchgate.net/figure/Typical-CBR-life-cycle-comprising-four-stages-Each-of-the-steps-comprising-the-CBR-life_fig1_221916036
- [6] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*, 3rd Edition. Pearson Education, Inc. Pearson Prentice Hall, 2008.
- [7] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*, 2nd Edition. Morgan Kaufmann Publishers, 2006.
- [8] <https://study.com/academy/lesson/what-is-the-median-definition-lesson-quiz.html>
- [9] <http://caia.swin.edu.au/urp/diffuse/ml.html>
- [10] <https://ieeexplore.ieee.org/document/1490524/>
- [11] <http://atvs.ii.uam.es/mcyt75so.html> (ATVS - Biometric Recognition Group >> Databases >> MCYT - SignatureOff - 75).
- [12] Amir Soleimani, Kazim Fouladi, and Babak N. Araabi. UtSig: A Persian offline signature dataset. *Institute of Engineering and Technology Biometrics*, 6(2017):1-8, 2017
- [13] Kai Huang and Hong Yan. Off-line signature verification based on geometric features extraction and neural network classification. *Pattern Recognition*, 30(1):9-17, 1997
- [14] H Baltzakis and N Papamarkos. A new signature verification technique based on a two-stage neural network classifier. *Engineering applications of Artificial Intelligence*, 14(1):95-103, 2001
- [15] Alan McCabe, Jarrod Trevathan and Wayne Read. Neural network-based handwritten signature verification. *Journal of Computers*, 3:9-22, 2008