International Journal of Rough Sets and Data Analysis
Volume 7 • Issue 1 • January-March 2021

Feature Engineering Techniques to Improve Identification Accuracy for Offline Signature Case-Bases

Shisna Sanyal, Jadavpur University, India
Anindta Desarkar, Jadavpur University, India
Uttam Kumar Das, Tata Consultancy Services, India
Chitrita Chaudhuri, Jadavpur University, India

ABSTRACT

Handwritten signatures have been widely acclaimed for personal identification viability in educated human society. But, the astronomical growth of population in recent years warrant developing mechanized systems to remove the tedium and bias associated with manual checking. Here the proposed system performs identification with nearest neighbor matching between offline signature images which are collected temporally. The raw images and their extracted features are preserved using case-based reasoning and feature engineering principles. Image patterns are captured through standard global and local features, along with some profitable indigenously developed features. Outlier feature values, on detection, are automatically replaced by their nearest statistically determined limit values. Search space reduction possibilities within the case base are probed on a few selected key features, applying hierarchical clustering and dendrogram representation. Signature identification accuracy is found promising when compared with other machine learning techniques and a few existing well-known approaches.

KEYWORDS

Case-Based Reasoning, Dendrogram, Feature Engineering, Hierarchical Clustering, Identification Accuracy, Image Case-Base, Nearest Neighbor Classification, Outlier Detection

1. INTRODUCTION

From ancient times, handwritten signature is the most well-known biometric characteristic for appropriate identification of a person or to authenticate a document. Biometrics is broadly categorized into two sections: behavioral and physiological. Handwritten signatures belong to the first category. The mode of collection is also the easiest and cheapest. For these reasons, it has been one of the most popular techniques favored so far. From financial and business transactions to invigilation in examination hall, this mode of identification has been used everywhere most profusely. Besides authenticating one's identity, other application areas include transaction confirmation, civil law contracts, acts of volition, personal cards, administrative forms, formal agreements, acknowledgement of received services etc. The wide and huge usage area obviously demand automatic identification

DOI: 10.4018/IJRSDA.20210101.oa1

technique, as in the present era of data avalanche, it would require abhorrent amounts of time and effort, if performed manually.

Acquiring training data is a continuous process – the signatures collected would usually come from different form-filling sessions. Signature of a person tends to changes with time – over and above the fact that no two signatures of a person are exactly the same. Hence the proposed approach has been made adequately robust and efficient to support the need of a heterogeneous and dynamic environment.

According to the data acquisition mechanism, signature identification system is classified into two types: online and offline. Online method needs special set of devices and instruments to capture the pen movements and pressure over the digital medium at the time of writing, thus involving sophisticated and costly tools. On the other hand, the offline technique at most needs a scanner or a digital camera at the input end to receive a digital representation of the signature in the pixel form corresponding to the grey level intensity at each point within the signature image. In this research work, external media such as a piece of paper have been used to capture the signature, which has thereafter been scanned to receive its digitized copy prior to storing it within the system. However, recent mobile devices allow easy capturing of online signatures, which can well be utilized to accommodate dynamic update of training set and test set data.

The objective of our research is to build a classifier which helps to detect the identity of a person, where training begins by comparing presented signature with each case or person preserved in the base. In an ideal situation, the training process is supposed to be improved if a prior clustering process partitions the case base into further segments, depending on some key feature values, leading to reduced searching and comparison times. For this purpose, a hierarchical clustering technique, using dendrograms, have been examined here to survey the prospects. The feature values extracted from the images have been discretized and the prospect of capping them within normal limits to exclude abnormal values, have been studied separately. Avoiding outlier values have been experimentally found to improve accuracy of identification, which has been one of the primary motivation behind the present work.

In our proposed work, Case Based Reasoning (CBR) technique is deployed to utilize incremental learning procedure at the beginning. The machine is trained, on sample signatures of each person preserved within a case base, which are utilized to recognize a particular person whose authentic signature is posed as a problem to be solved by the system. As a further motivating factor, it may also be mentioned here that, as the literature survey in the following section reflects, no comparative work in the domain utilizes CBR methodologies. This novel approach, of utilizing CBR techniques with straight-forward feature engineering in the form of manipulative outlier handling, has been truly justified by the accuracy results obtained with an indigenous dataset accumulated by the present researchers. Here signatures are stored as the combination of various attributes fetched from the captured images. In the domain of machine learning, features play a very important role, as recognition primarily depends on them. The terms attribute and feature shall be used interchangeably in the present document. These collected features may be considered to be the part of a set problem, the solution to which is the identity of the signatory attached as the class value. Once the newly posed part of a problem is matched sufficiently with an already preserved part of a problem in a case, the solution part of the case is the required output in the form of personal identity. The significance of proper Feature Engineering, and Outlier Handling therein, is thus self-evident under the circumstance.

Following Section 2 highlights a few researches on signature identification. Section 3 describes some basic concepts associated with the proposed system. The proposed techniques for identification are presented in Section 4. It also includes the detailed description of datasets and system tool configurations. Experimental results are depicted in Section 5. The final section concludes with inferences drawn from the results, as well as some future directions of research which may further augment the accuracy and performance of the system.

2. RELATED WORK

Diaz et al. (Diaz, Impedovodo, Malik, Pirlo & Plamondon, 2019) presents an update on automatic signature identification and verification technologies practiced in the last decade. They report that, though systems have traditionally been developed largely to cope with English-based signatures, they have paid attention to the growing use of other scripts, such as Arabic, Persian, and Chinese. They also point out problems associated with assessment of the consent form at the acquisition phase, and that of acquiring license agreement for research dataset releases. This invaluable document also carries a list of the popular Handwritten Signature Databases publicly available including the last two datasets used in the currently proposed work.

A signature identification as well as verification system using Contourlet Transform (CT), developed by Pourshahabi et al. (Pourshahabi, Sigari & Pourreza, 2009) propose noise removal of signature images followed by size normalization. The identification accuracy results claimed with two sets of data, one a Persian signature base for 20 persons and the other an English signature base of 22 persons, are 100% (too perfect in our opinion) and 93.2% respectively . Although both datasets are meagre in size, the second one involving the Stellenbosch dataset has been used to benchmark the proposed system at the experimental stage.

The above mentioned dataset derives its name from the Stellenbosch University, where an automatic authentication system for offline handwritten signatures was developed by Coetzer et al. (Coetzer, Herbst & du Preez, 2004) using the discrete Radon transform (DRT) and a hidden Markov model (HMM) on the same dataset. It consists of multiple authentic signatures, collected over a spread of two weeks, from each new entrant to the University. The signatures were subsequently digitized at 300 dpi resolution using a flatbed scanner and the images are claimed to be adequately preprocessed to make them noise-free. The set can be retrieved from the university website on request.

Another identification-cum-verification technique proposed by Kalera et al. (Kalera, Srihari & Xu, 2004) offer for offline signatures applying quasi-multi-resolution procedure for extracting GSC (Gradient, Structural and Concavity) features. Results reported lie in the range 93.18% as the identification accuracy. The technique involved testing statistical distance distributions between different writers, and that within one's own set. The nomenclature used for the dataset by CEDAR is Database A. It consists of 24 genuine offline signature samples collected from each of 55 participants within a pre-assigned 2"x2" writing space. These samples are also scanned at 300 dpi to produce 8-bit gray-scale PNG format images.

A separate group of researchers, Fotak et al. (Fotak, Bača & Koruga, 2011) have presented a promising approach to identify handwritten signatures by constructing signature graphs, based on the number of strokes produced by the writer. Graph norms are then utilized to classify valid users of an authentication system, thereafter comparing each new signature graph with values stored in the validated system base. A deep learning model for offline handwritten signature identification followed by verification is described by Ribeiro et al. (Ribeiro, Gonçalves, Santos & Kovacec, 2011). The architecture is modeled on representational layers presented by human mental ability. Advanced features such as Tri-Surface, Six-Fold Surface, Best Fit, as well as K-Means, Histogram of frequencies, Discrete Cosine Transform and Wavelet Transform are utilized, together with Modified Direction Feature, and standard features such as Width, Height, and Geometric Parameters (Polar and Cartesian).

An innovative approach, Adaptive Window Positioning technique has been utilized successfully for accurate signature feature extraction to identify the offline handwritten signatures by Sulong et al. (Sulong, Ebrahim & Jehanzeb, 2014). This process claims to have the ability to verify and identify of an individual's signature even if produced under emotional stress. Elhoseny et al. (Elhoseny, Nabil, Hassanien & Oliva, 2018) presents a system utilizing rough neural network and rough set, to achieve better recognition performance. Rough sets provide a range of techniques to organize data at concept level, which in turn helps at classification as well as analysis based on such data. The hybrid system

not only aids in data cleaning and overall pre-processing, but also achieves success in disclosing hidden patterns and meanings by employing better features of both technologies.

Another approach, integrating the methods of image processing with machine learning techniques such as neural networks as well as statistical methods, is discussed by Mohammed (2019). Here a Gabor filter using multi-scale and multi-orientation analysis methods distinguishes the different signatures, mapping them to their true owners to establish identity. The images converted to the frequency space using Gabor converter are then utilized to extract their statistical and engineering characteristics, ultimately to form a matrix to be used as the input to a back-propagation neural networks. The signature images were collected from people associated with the University of Mosul. Reported success rate is claimed to be 88.57%.

3. BACKGROUND

3.1 Machine Learning Techniques

The field of Machine learning is an integral part of Artificial Intelligence (AI). Machine Learning tools help a system to accept the input and improve at its task automatically on its own without further programming effort, discussed by Mitchell (1997). Algorithms pertaining to machine learning belong to two basic types – Supervised and Unsupervised strategies, presented in the book authored by Bishop (2006).

In the present work, instance based classification techniques are utilized to identify a person by comparing distances between feature vectors representing offline handwritten signature images of that person. Clustering procedures amongst similar signatures are also tried out for expected reduction in search times.

3.1.1 Hierarchical Clustering Using Dendrogram

Clustering involves finding a structure in a collection of unlabeled data, leading to discovery of a new set of categories intrinsically, presented in the research by Nirmala et al. (Nirmala & Saravanan, 2014). Hierarchical clustering is an approach to identify groups in the dataset which does not require specifying the number of clusters formed. The groups are nested and organized as a tree. The trees can be formed either using a Top-to–bottom (Divisive) approach or a Bottom-Up (Agglomerative) approach using a special structure known as Dendrogram. The technique adopted in this research utilizes the Agglomerative approach.

Here, at the beginning, each data tuple is considered to be in a separate cluster and this is the bottom-most level. Clusters are merged based on the degree of similarity from the next level onwards, until at the topmost level there appears a single cluster containing all the data tuples. The dendrogram thus formed needs to be split vertically at the correct level using pre-defined stopping criteria, mentioned in the research by Murtagh (1983).

Dendrograms are used to illustrate the arrangement of the clusters produced by the above mentioned techniques. Our research has adopted this technique where the training dataset can be divided into clusters depending upon the linkage and distance methods used. The incoming test signature is matched with the index of each of the corresponding clusters and the best matching cluster is chosen for further consideration. The chosen cluster is searched exhaustively for the best match.

3.1.2 Case Based Reasoning (CBR)

Classification, a supervised learning process, is a form of data analysis which can be used to extract models describing important data classes or to predict future data trends. Data classification, in general, is a two-stage process. At first, the above-referred model is built describing a predetermined set of data classes. The model is trained by feeding it with database tuples having known class values. Beside the class value, each tuple is composed of discerning attribute values. Such data tuples are

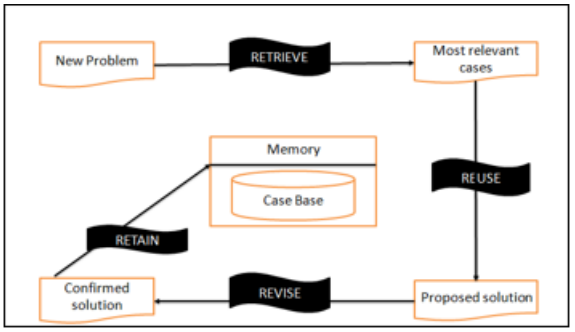
also referred to as samples or objects and are collectively known as the training set. The second stage involves testing the model with a set of samples, known as the testing set, which determines the performance accuracy of the model. Following parts describe some classification techniques adopted in the current work.

Case Based Reasoning or CBR is a lazy learning instance-based classification technique that treats every problem-solution pair as a case and stores each such case in a base. The laziness of the classifier lies in its reluctance to build a learning model at the beginning. It learns and classifies as and when it gets fresh instances, which is another human-like aptitude that CBR shows. The term 'lazy' is used here in contrast to the group of 'eager' learners such as ANN, Decision Tree, or SVM – all of which build classifier models in the training phase. In fact, it gains in terms of time as it does not need to build a model apriori! Further, each unsolved problem is supplemented with its correct solution which represents its class value. Thus, Case Base is a storage area where every case is a combination of a detailed statement of the problem and its solution, along with the necessary metadata required for that problem. The process of insertion of a new case into the case base is preceded by search among the old cases in the base with the help of some predesigned indexing system. If an exact match is found there is no need to insert the new case. Otherwise, some nearest matches are found whose class information help to provide the new case with the suitable class. The new case with the new solution is now ready for insertion into the case base; the strategy is discussed in detail by (Riesbeck & Schank, 2013) and amply supported by (Aamodt & Plaza, 1994; Aha, 1998) in their researches in this domain.

So CBR is a problem-solving strategy and an AI technique that considers previous cases to take decision for new situations. The previous cases are constituted of past experience rather than rules, and CBR work by recalling similar cases to find solution to new problems. The CBR process includes four main steps:

RETRIEVE similar cases from the case base for a targeted problem.
If the retrieved problem matches the current problem exactly, **REUSE** its solution part.
Otherwise, **REVISE** retrieved solutions to produce new solution, and lastly
RETAIN the new problem solution pair in the case base.
The process is depicted in the following Figure 1.

Figure 1. Life cycle of a CBR based system



The advantages of using CBR technology are manifold. First of all, the case-base itself provides a structure for retaining metadata as well as actual data along with class information per tuple, which is particularly useful for problem solving in domains utilizing complex symbolic descriptions. Secondly, the effort of classification is greatly reduced if an exactly matching case already exists within the base, when only the solution part or class value is returned in answer to a query. This is

very similar to the human reasoning process, where the experience of an expert in the area is directly applied. Further, properly introduced indexing techniques enhances the effort of finding relevant and similar tuples efficiently. Finally, the CBR allows incremental learning procedure, again a human-like feature, which helps the knowledge base to improve its expertise of prediction in terms of time, efficiency and accuracy.

3.1.3 Nearest Neighbor Classification Technique

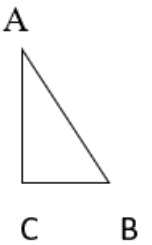
The Nearest neighbor approach allows the user to retrieve cases based on a weighted sum of features in the input cases that match the cases in memory. Every feature in the input case is matched to its corresponding feature in the stored or old cases and the degree of match for each pair is computed. One of the most obvious measures of similarity between two cases is the distance, as described by Duda et al. (Duda, Hart & Stork, 2012).

In the current research work, Nearest-neighbor classification technique is used to retrieve similar cases or training samples, based on least distance measured with respect to the accumulated difference of attribute values between the test and training samples.

3.2 Distance Measurement

Distance is the length of space between two points. In our identification program, we treat each signature as a point in a n-dimensional space represented by a vector, where each element in the vector is represented by a feature value, courtesy Han et al. (Han, Pei & Kamber, 2011). So to calculate the similarity (or difference) between two signature images, we can compare the utility of several such distance measures. We have chosen here two distance measures, namely Euclidean distance and Manhattan distance. The concept is based on the geometric distances calculated between two points connected by the hypotenuse of a right angled triangle drawn in a two-dimensional plane as depicted in Figure 2 below.

Figure 2. Right Angled triangle ABC in a 2-dimensional plane



The **Euclidean** and **Manhattan Distance** measures are obtained from Fig. 2 above – the first one depicts the Hypotenuse AB and the second one the sum of Perpendicular AC and Base CB.

3.3 Image Processing

In this research work, we are working on offline signature images. Image processing deals with methods which convert an image into digital form and operates on it, in order to get an enhanced image or to extract some useful information from it. The term 'Image processing' is loosely used to refer to Digital Image Processing in general. The present work involves mostly techniques associated with digital images, as explained in the excellent book by Gonzalez et al. (Gonzalez & Woods, 2008) on this topic. Image samples usually have lots of noise. Some parts may appear to be blurred, some

others may have non-uniform borders, and some are not properly aligned due to technical problems with the scanner, or issues associated with the pen or the paper. Most procured images thus need to be cleaned and pre-processed. The different pre-processing steps employed here include: Image Binarization, Noise Reduction, Image Cropping, Skeletonization and Image Resize.

3.4 Attributes and Their Usages

In a computing environment, an attribute defines a specific property of entity according to Han et al. (Han, Pei & Kamber, 2011). The different types of attributes required for describing offline handwritten signature images can be broadly classified into two categories - the global attributes and local attributes. Attributes collected from the signature as a whole are termed as global attributes, such as the number of black pixels in the complete binarized image. On the other hand, the attributes constructed from one part of the signature image are known as local attributes, for example angular distances from pixels within each vertical quarter of the image.

Most of the global attributes are acquired automatically at the basic image level. Some of the indigenously developed attributes, such as the local angular distance features and reduced component features are obtained at a higher level integration. The global attributes used in this context are a popular set, providing efficient classification results as high as 93.8% in the domain of signature identification, shown in their research by Chaudhuri et al. (Chaudhuri & Chaudhuri, 2016). The indigenously developed local attribute set has been incorporated as it has displayed better authentication abilities for signatures as reported by Desarkar et al. (Desarkar, Sanyal, Baidya, Das & Chaudhuri, 2019). The details of all these attributes are provided in the next section.

4. SYSTEM METHODOLOGY

4.1 Proposed Attributes

Manual processes of checking signature entail the chance of personal bias and inefficiency. To utilize the machine for removing such bias and inefficiency we need to present the signature to the machine either as a digital image or as a sequence of features extracted from such an image. Here we have taken advantage of both procedures and extracted two major sets of attributes for our system.

4.1.1 Local Attributes

Local attributes here comprise of 40 angular distance features, presented and indigenously developed by the internal research team, Desarkar et al. (Desarkar, Sanyal, Baidya, Das & Chaudhuri, 2019). They are termed as angular distance features, as they represent some angular i.e., radial distance from centre of mass and are constructed according to the following algorithm.

4.1.2 Global Attributes

Features related to the structure of the signature image as a whole are categorized as global. These global features are usually extracted from the pixels that lie within the region circumscribing the signature image. The global features are easily extractable and have been used alone and along with the earlier described local features in our classification model. A complete list of these features are given below in Table 1, some extracted from standard papers presented by (Baltzakis & Papamarkos, 2001; Huang & Yan, 1997; McCabe, Trevathan & Read, 2008) and others (marked with *) indigenously developed as already mentioned in the earlier section.

4.2 Proposed Techniques For Identification

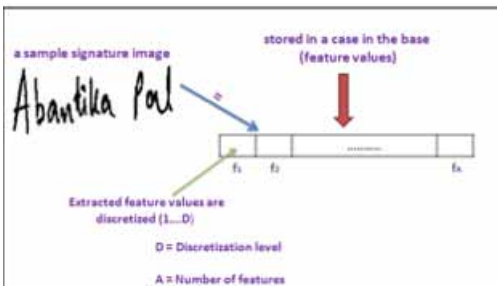
In this technique, as already mentioned, CBR methodology is utilized to establish a person's identity. Here two case bases are being used. The first amongst these is named the Person case base, where

Algorithm: Angular-Distance-Feature
Input :- Pre-processed Signature Image
Output:- 40 Angular Distance Local Features
Method:-
1. Image re-sized to 400x100 pixel-size image
2. Image sub-divided into 4 equal vertical parts
3. For each sub-part of size 100x100 pixels
4. do
5. Calculate centre of mass for the part
6. Calculate the geometric centre of the part.
7. Calculate distance d1 from centre of mass to geometric centre of part
8. Calculate angle a1 subtended by line joining these two points
9. Divide part into 8 angular regions at 45° interval around centre of mass
10. For each of these equal angular regions
11. Find farthest black pixel
12. If black pixel found
13. Compute Euclidean distance d of black pixel from centre of mass
14. Else
15. Set distance d to 0
16. EndFor each angular region
17. EndFor each part
18. Map all d values into the first 32 (=4x8) angular distance features
19. Map the <d1, a1> pair for each part into the last 8 (=4x2) feature values

Table 1. List of Global Attributes

Signature Height	**Top heaviness**	**Interior to Exterior Pixel Ratio ***
Pure Width	Baseline Shift	Number of significant component
Image Area	Horizontal Dispersion	Reduced Component *
Vertical Centre	Number of edge points	Maximum Vertical Projection
Horizontal Centre	Number of cross points	Maximum Horizontal Projection
Global Slant	Number of closed loops	Vertical Projection Peaks
Local Slant	Mean Ascender Height	Horizontal Projection Peaks
Mean Slant	Mean Descender Depth	

Figure 3. Sample Signature Structure within Person Case Base



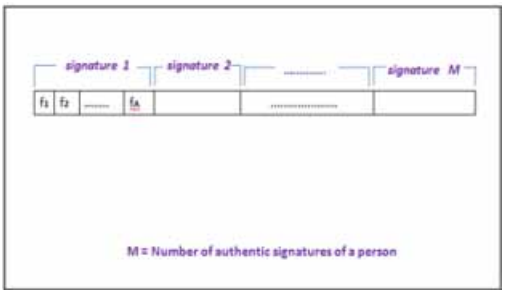
we preserve authentic signature list of each person and their general information. The second one is called the Feature case base.

4.2.1 Person Case Base

The structure of a single signature within a case from the Person case base is depicted in Figure 3 below. Here the features extracted from the pre-processed image are discretized on the basis of a preset level. These feature values are then stored as part of the case.

The set of training signatures, which is housed as a recorded problem within a case, is a composition of feature list as shown in Figure 4 below. The sample signatures collected at the outset undergo a dynamic screening process. Ultimately the best and latest ones are preserved with an eye to make the system sufficiently robust and up-to-date, as presented in an earlier research work by Chaudhuri et al. (Chaudhuri & Chaudhuri, 2016)

Figure 4. Typical Training-set Signatures for a person

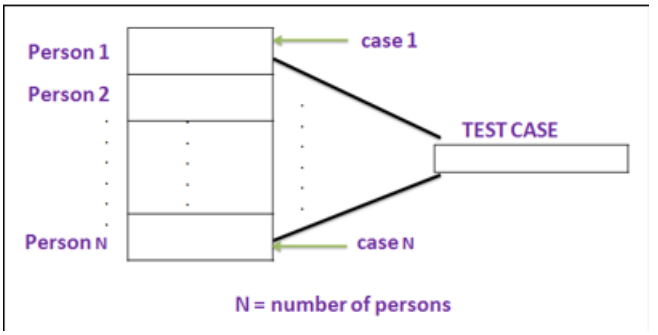


The overall process of matching is depicted in Figure 5 below. As described above, the Person case base already houses sample training signature sets as the authentic problem part for each and every candidate person, along with the identity of that person kept in the solution part. When a new test signature arrives, it is compared with each of the cases, and the nearest match(/es) collected. The collection produces the solution as the predicted identity of the author of the test signature.

4.2.2 Feature Case Base

The Feature Case Base houses the statistical information for each and every feature extracted from the signatures as depicted in figure 6 below. These ultimately help to produce the discretized feature value as described in the following algorithm. We calculate the three quartile values Q1, Q2 and Q3

Figure 5. Process of Matching Cases



from the total range of values for each feature. Next, the Inter quartile range IQR is calculated to detect outlier feature values. For forcefully maintaining all feature values within the allowed tolerable range, we need to calculate the Tolerable Lower limit TL and Tolerable Upper limit TU, whenever we apply capping to our feature values.

Algorithm : Calculate Discretized Attribute Value

Input : Attribute Value A
Discretized Level D
Output: Discretized Attribute Value DAV

Method :
IQR= Q3-Q1
TL= Q1-1.5 x IQR
TU= Q3-1.5 x IQR
If capping applied
Then
    If A < TL
    Then A= TL
    Else
        If A > TU
        Then A= TU
        EndIf
    EndIf
EndIf
DAV= ((A-TL)/(TU-TL)) x D

In this scenario, two approaches to the Identification process are introduced – the first one involving capping of attributes and the second one involving no capping of attributes depicted as schematic flow diagrams in Figures 7 and 8 below.

The following figure depicts the identification process adopted by the present system. On a trial basis, we at first utilized a hierarchical clustering technique based on dendrograms to partition the signature case-base into smaller units to facilitate faster identification.

Further experiments revealed better results when the case-base was considered in totality. In either case, the Classifier predicts the identity of the test sample based on its match with some case value, indicated by the smallest distance obtained after comparison with the case base, searched either partially [Figure 9] or entirely. So, to avoid deterioration of identification accuracy, we reverted back

Figure 6. Feature Case Base

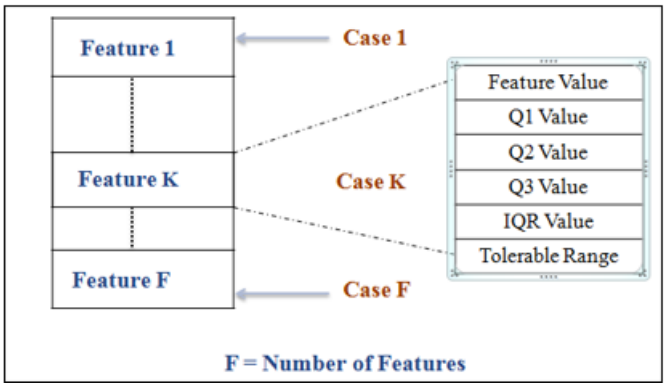


Figure 7. Identification with capped attribute: schematic flow diagram

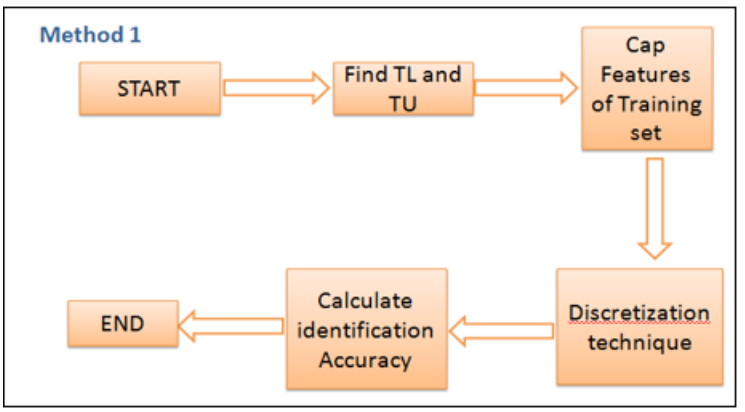


Figure 8. Identification without capping: schematic flow diagram

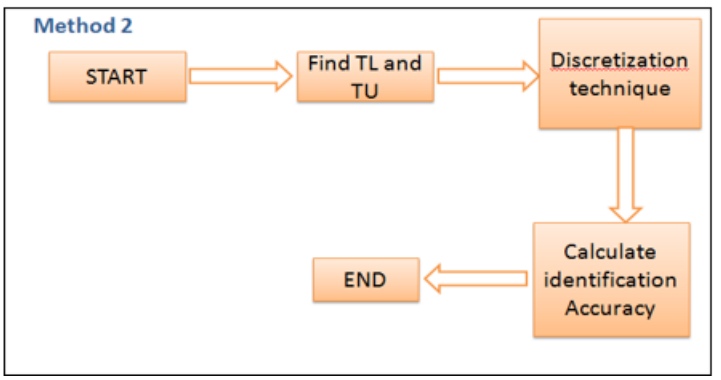


Figure 9. Identification Process utilizing Signature Case Base

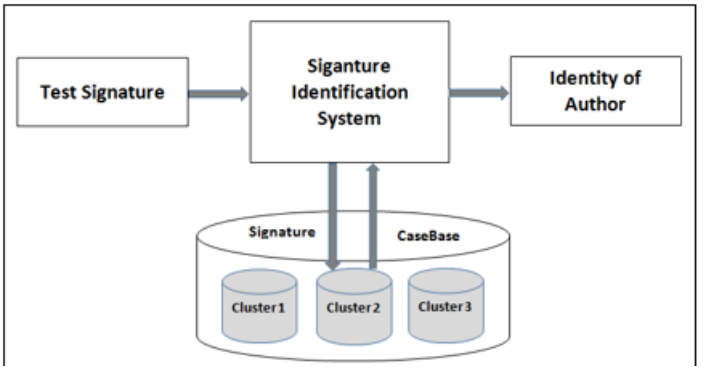
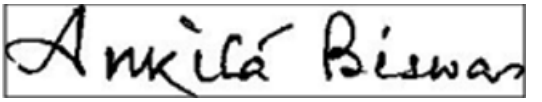


Figure 10. Sample Signature of Dataset 1



to the original scheme of considering the signature case base as a whole. The results of both trials have been reported under the section headed Results.

4.3 Source Datasets

Three Datasets of images have been used in the experimental stage – designated as Dataset 1, Dataset 2 and Dataset 3. A detailed description of each follows.

Dataset 1: This is an indigenous dataset, comprising of signatures collected within Jadavpur University campus, mainly from student volunteers and faculty members. 121 persons submitted 20 authentic signatures each. 20 more forged signatures were prepared, by available student artists, producing 10 skilled or exact reproductions and 10 random or arbitrary signatures for each of these 121 persons. The total tally thus comes to 4840 signatures overall. The signatures were scanned at a resolution of 200 dpi and stored in the PNG (Portable Network Graphics) format. For identification purpose we are considering only the authentic signatures. Amongst the 20 authentic signatures 15 are used for training and 5 for test – a ratio of 3:1. Following Figure 10 is a sample signature from Dataset 1.

Dataset 2: The attributes in this dataset have been obtained from the standard MCYT Bimodal Bio-metric Database (Diaz et al, 2019; Ortega-Garcia et al., 2003). Here there are 2250 signature images of 15 authentic and 15 skilled forged signatures for each of 75 persons. Among the 15

Figure 11. Sample signature of Dataset 2

Figure 12. Sample Signature of Dataset 3

authentic signatures of each person, 10 signatures were segregated for training and the remaining 5 were utilized to test the system maintaining the thumb rule of two thirds and one third standard ratio. All images were reportedly scanned at a resolution of 300 dpi. Following Figure 11 is a sample signature from Dataset 2.

Dataset 3: In this standard Dataset 3105 signatures are said to have been collected from graduate students of University of Tehran and Sharif University of Technology by Soleimani et al. (Soleimani, Fouladi & Araabi, 2016). Here there are 27 authentic signatures per person. 20 of the authentic signatures belonging to each of the 115 participants were utilized as the training dataset, while the remaining 7 signatures per person were reserved as the test dataset – again approximately a 3:1 ratio. The images were mentioned to have been scanned at 600 dpi resolution and stored in the PNG (Portable Network Graphics) format. Following Figure 12 is a sample signature from Dataset 3.

4.4 System Tools

The techniques used in this research work have been developed using two standard tools

- Matlab R2017a
- Python 3.5.

5. RESULTS AND PERFORMANCE ANALYSIS

Three datasets are used in the experiment where the major part in each set are utilized for model building and rest of the set is treated as test dataset on which signature identification accuracy is calculated.

5.1 Result-Set Using Hierarchical Clustering

As mentioned earlier, hierarchical clustering is applied in our research to reduce search space and time complexity. The two global features Width and Interior to Exterior Pixel Ratio have been used here to form three clusters within the compact Signature Database using dendrogram technology as shown in Figure 13 below.

The following Table 2 specifies the identification accuracy percentages obtained on the fragmented signature base using the above method.

However as the accuracy percentages reflected in the experiments have not yet been of a promising nature, further studies have been carried out on the complete case base, as already mentioned.

5.2 Result-Set Without Applying Clustering

For comparing the test signature with the signatures within the case base, we have utilized the well-known Nearest-Neighbor Classification Techniques based on the two distance measures, Euclidean and Manhattan, as depicted in Table 3 below.

Figure 13. Dendrogram

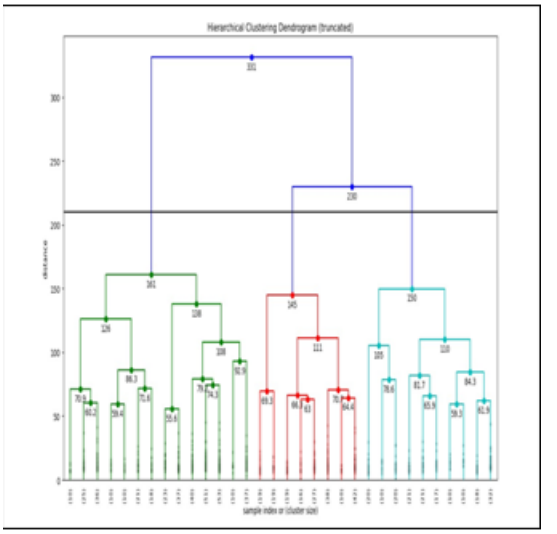


Table 2. Identification Accuracy for Clustered Signatures

Attribute Type	Dataset	Accuracy %
Global	1	74.67
	2	62.13
Local	1	73.6
	2	60.53

Table 3. Identification Accuracy for Unclustered Signatures

Dataset	No. of Person / Cases	Method	Accuracy % (W/O Capping)	Accuracy % (With Capping)
1	121	Euclidian	94.8760	94.7107
		Manhattan	97.1901	**97.1901**
2	75	Euclidian	72.5333	72.5333
		Manhattan	81.3333	**82.6667**
3	115	Euclidian	71.6770	73.2919
		Manhattan	83.4783	**87.5776**

Figure 14. Histograms for datasets using Manhattan distance

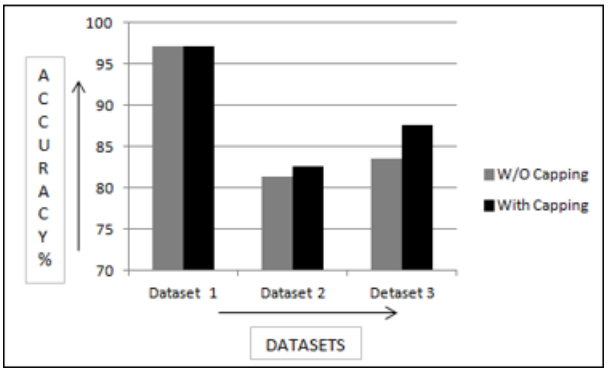


Table 4. Accuracy Percentage of Different Classifiers for Datasets 1 and 2

Classifier Model	Accuracy % - Dataset 1	Accuracy % - Dataset 2
CBR Manhattan (capped)	**97.19**	82.67
CBR with DTW	93.8	82.13
MLP	96.66	88
SVM	96.53	89.6
NB	72	73.33
DT	84.27	81.07
KNN	80.53	65.87

Table 5. Accuracy Percentage – Proposed System vs Related Works

Reference	Dataset Name	Dataset Characteristics	Sample/Case Count	Identification Accuracy
Proposed	**Dataset 1**	PNG Format	**2420**	**97.19%**
Pourshahabi et al., 2009	Stellenbosch	PNG Format	660	93.20%
Kalera et al., 2004	Database A	PNG Format	1320	93.18%

The above table clearly indicates that Manhattan distance using capped value method provide the best accuracy for each of the datasets. So, in the following Figure 14, the accuracy results using Manhattan distance measures only have been pictorially represented in the form of vertical histograms for all three dataset.

To establish the usefulness of the present CBR system using capped Manhattan distances, its performance have been compared with an earlier CBR system (Chaudhuri & Chaudhuri, 2016) designed utilizing Dynamic Time Warping distance between two authentic signatures of the same person, along with the accuracy levels of several other standard classifiers, all performing on the first two sets of data. Amongst these other classifiers the Multi-Layer Perceptron (MLP) model is built with a single hidden layer and based on a sigmoidal activation function, using the WEKA Data

Mining Software, according to theories discussed in the classical literature on the topic by Haykin (2004).The remaining classifiers include the Support Vector Machine (SVM), the Naïve Bayes (NB), the Decision Tree (DT) and the K-Nearest-Neighbor (KNN) model, all developed using the MATLAB tool. The SVM used linear kernel and sequential minimal optimization technique for separating the hyperplane (Vapnik, 1998; Vapnik, 2013).The NB was designed having Normal(Gaussian) Distribution with uniform prior (Duda, Hart & Stork, 2012). The split criterion for DT was based on Gini's diversity index (Breiman, Friedman, Stone & Olshen, 1984), and the KNN model was built for k = 3 (Duda, Hart & Stork, 2012). The comparison results for the first two datasets are presented in the following table 4.

In the final stage, the performance of the proposed system is compared against two previous techniques and the results tabulated in the following Table 5.

The supremacy of our proposed system is favorably established, considering both the dataset size as well as the accuracy percentage achieved. The details of statistical significance of the system is available in the research by Desarkar et al. (Desarkar, Sanyal, Baidya, Das & Chaudhuri, 2019).

6. CONCLUSION AND FUTURE SCOPE

Identification of the author of a handwritten offline signature is achieved in this research work through deploying the instance-based classification technique associated with Case Based Reasoning. General improvement affected by capping feature values within tolerable range, and that too with the Manhattan distance measure, has been comprehensibly demonstrated in results tabulated in Table 3 of the above section and the corresponding graph recorded in the following Figure 14. The relatively better recognition rate achieved overall by the indigenous Dataset 1 may partially be contributed by the higher training to testing set ratio for it, compared to Dataset 2. Dataset 3 ranks second in this respect, in spite of its script pattern complexity and diversity. The subsequent Table 4 enumerates a clear ascendancy of the proposed CBR system over other Machine learning techniques, when augmented by the capped manhattan distance comparison technique, in predicting the correct identity for the indigenous Dataset 1. Table 5 in the same section also reflects this success rate of 97.19% for the said dataset, which is substantially higher than that achieved by either of the other related works reported in (Pourshahabi, Sigari & Pourreza, 2009) and (Kalera, Srihari & Xu, 2004).

The case-base utilized here serves several purposes. Preliminarily, it allows storing sets of authentic signatures of each candidate person, both as raw / pre-processed image, as well as in the form of machine decipherable attribute value collections for each signature. The second utility of a case, preserved for each individual, is including personalized information such as discrepancies or deviations caused by normal ageing process or accidental infirmities, which may be required later to bias the decision-making process intentionally. The third advantage is achieved by the incremental learning feature inherently adopted by a CBR system – here it allows updation of the case-base with more recent specimens through usage of proper maintenance of sequence within the system as and when required. This aspect is captured by a dynamic time stamping process associated with the collected signatures within our indigenous dataset as already explained by Chaudhuri et al. (Chaudhuri & Chaudhuri, 2016) in their earlier research.

Case-bases naturally lend themselves towards fast indexing techniques to reduce searching time for partial or exact case-matches. We sought to achieve faster access by partitioning the case-base into smaller clusters relative to indices formed from some choice attribute values and applying hierarchical clustering techniques using dendrograms. But in spite of its sound theoretical validity, the tactics failed to live upto its promise as made apparent in the tabulated result for the clustered signature case-bases. However, further trials with different attribute combinations may lead to better success levels in future works. There remain a number of unexplored grounds within the research space – such as trying out other forms of distance measures such as Jaccard and Cosine indices between attribute values - and seeking a more complete performance measure by indicating precision, recall and F1-score. These

could not be accommodated in the present work due to paucity of time and space, but must remain as lucrative essentials in furthering this endeavor.

REFERENCES

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, *7*(1), 39–59. doi:10.3233/AIC-1994-7104

Aha, D. W. (1998). The omnipresence of case-based reasoning in science and application. *Knowledge-Based Systems*, *11*(5-6), 261–273. doi:10.1016/S0950-7051(98)00066-5

Baltzakis, H., & Papamarkos, N. (2001). A new signature verification technique based on a two-stage neural network classifier. *Engineering Applications of Artificial Intelligence*, *14*(1), 95–103. doi:10.1016/S0952-1976(00)00064-6

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Chaudhuri, C., & Chaudhuri, A. (2016). Auto-Upgrading of Signature Case-Base used for Efficient Detection of Identity and Authenticity. *International Journal of Research in Engineering and Applied Sciences*, *6*(5), 157–172.

Coetzer, J., Herbst, B. M., & du Preez, J. A. (2004). Offline signature verification using the discrete radon transform and a hidden Markov model. *EURASIP Journal on Advances in Signal Processing*, *2004*(4), 925026. doi:10.1155/S1110865704309042

Desarkar, A., Sanyal, S., Baidya, A., Das, A., & Chaudhuri, C. (2019). Innovative Outlier Removal Techniques to Enhance Signature Authentication Accuracy for Smart Society. *International Journal of Distributed Systems and Technologies*, *10*(2), 64–83. doi:10.4018/IJDST.2019040104

Diaz, M., Ferrer, M. A., Impedovodo, D., Malik, M. I., Pirlo, G., & Plamondon, R. A. (2019). Prespective Analysis of Handwritten Signature Technology. *AMC Computational Survey, 51*(6).

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Elhoseny, M., Nabil, A., Hassanien, A. E., & Oliva, D. (2018). Hybrid rough neural network model for signature recognition. In *Advances in Soft Computing and Machine Learning in Image Processing* (pp. 295–318). Springer. doi:10.1007/978-3-319-63754-9_14

Fotak, T., Bača, M., & Koruga, P. (2011). Handwritten signature identification using basic concepts of graph theory. *WSEAS Transactions on Signal Processing*, *7*, 117–129.

Gonzalez, R., & Woods, R. (2008). *Digital Image Processing* (3rd ed.). Pearson Prentice Hall.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Haykin, S. (1999). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Huang, K., & Yan, H. (1997). 'Off-line signature verification based on geometric feature extraction and neural network classification. *Pattern Recognition*, *30*(1), 9–17. doi:10.1016/S0031-3203(96)00063-5

Kalera, M. K., Srihari, S., & Xu, A. (2004). Offline signature verification and identification using distance statistics. *International Journal of Pattern Recognition and Artificial Intelligence*, *18*(07), 1339–1360. doi:10.1142/S0218001404003630

McCabe, A., Trevathan, J., & Read, W. (2008). Neural network-based handwritten signature verification. *Journal of Computers*, *3*(8), 9–22. doi:10.4304/jcp.3.8.9-22

Mitchell, T. M. (1997). *Machine learning Machine Learning* (International Edition). McGraw-Hill.

Mohammed, I. S. (2019). Handwritten Signature Recognition with Gabor Filters and Neural Network. *Second International Conference of Mathematics (SICME2019), AIP Conf. Proc., 2096*. doi:10.1063/1.5097802

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, *26*(4), 354–359. doi:10.1093/comjnl/26.4.354

Nirmala, A. M., & Saravanan, S. (2014). A Study on Clustering Techniques on Matlab. *International Journal of Scientific Research*, *3*(11), 1497–1502.

Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez-Zanuy, M., Espinosa, V., & Escudero, D. et al. (2003). MCYT baseline corpus: A bimodal biometric database. *IEE Proceedings. Vision Image and Signal Processing*, *150*(6), 395–401. doi:10.1049/ip-vis:20031078

Pourshahabi, M. R., Sigari, M. H., & Pourreza, H. R. (2009, December). Offline handwritten signature identification and verification using contourlet transform. In *2009 International Conference of Soft Computing and Pattern Recognition* (pp. 670-673). IEEE. doi:10.1109/SoCPaR.2009.132

Ribeiro, B., Gonçalves, I., Santos, S., & Kovacec, A. (2011, November). Deep learning networks for off-line handwritten signature recognition. In *Iberoamerican Congress on Pattern Recognition* (pp. 523–532). Springer. doi:10.1007/978-3-642-25085-9_62

Riesbeck, C. K., & Schank, R. C. (2013). *Inside case-based reasoning*. Psychology Press. doi:10.4324/9780203781821

Soleimani, A., Fouladi, K., & Araabi, B. N. (2016). UTSig: A Persian offline signature dataset. *IET Biometrics*, *6*(1), 1–8. doi:10.1049/iet-bmt.2015.0058

Sulong, G., Ebrahim, A. Y., & Jehanzeb, M. (2014). Offline handwritten signature identification using adaptive window positioning techniques. *Signal and Image Processing: an International Journal*, *5*(3), 13–24. doi:10.5121/sipij.2014.5302

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag. doi:10.1007/978-1-4757-2440-0

Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons.

Shisna Sanyal has been awarded Master of Technology in Computer Science from Jadavpur University and is currently pursuing research under the supervision of Dr. Chitrita Chaudhuri in Jadavpur University in the domain of data mining.

Anindita Desarkar has been awarded Master of Technology in Computer Science in 2017 and is currently pursuing PhD under the supervision of Dr. Chitrita Chaudhuri and Dr. Ajanta Das from Jadavpur University in data mining domain. Her research area includes Evolutionary Algorithms, Data Mining and Data Analytics. She has already publications in International Journals, Conferences and Book Chapters. Anindita is having more than fifteen years of Industry experience in Data Warehouse domain in Tier 1 Company.

Uttam Kumar Das has been awarded his Bachelors degree in Mathematics (Honours) from Calcutta University and thereafter his Master of Computer Application degree from Jadavpur University in 2018. He is currently working as a System Engineer in Tata Consultancy Services in the Automation Testing domain. His research interest includes Image Processing, Data Analytics and Natural Language Processing. He pursued his final semester research work under the supervision of Dr. Chitrita Chaudhuri in Jadavpur University in the domain of data mining. He has contributed in International Conference papers as well. His hobbies include playing badminton, volleyball and organizing events. He has actively participated in the placement schemes and organized several cultural and technological events during his academic career in Jadavpur University.

Chitrita Chaudhuri is Associate Professor in the Department of Computer Science of Jadavpur University. She received Bachelor and Master degree in Electronics & Tele Communication Engineering from Jadavpur University in 1980 and 1982 consecutively. Prior to joining Jadavpur University she has also worked with Birla Institute of Technology, Mesra, Kolkata Extension Centre from 1998 to 2001. Dr. Chaudhuri has been awarded PhD in Computer Science from Jadavpur University in 2016. Her field of specialization include data mining and machine learning.