

# Efficient Search and Retrieval in Biometric Databases

Amit Mhatre, Srinivas Palla, Sharat Chikkerur and Venu Govindaraju  
Center for Unified Biometrics and Sensors, University at Buffalo, NY  
{ajmhatre,spalla2,ssc5,govind}@buffalo.edu

## ABSTRACT

Biometric identification has emerged as a reliable means of controlling access to both physical and virtual spaces. Fingerprints, face and voice biometrics are being increasingly used as alternatives to passwords, PINs and visual verification. In spite of the rapid proliferation of large-scale databases, the research has thus far been focused only on accuracy within small databases. In larger applications, response time and retrieval efficiency also become important in addition to accuracy. Unlike structured information such as text or numeric data that can be sorted, biometric data does not have any natural sorting order. Therefore indexing and binning of biometric databases represents a challenging problem. We present results using parallel combination of multiple biometrics to bin the database. Using hand geometry and signature features we show that the search space can be reduced to just 5% of the entire database.

**Keywords:** Binning, Hand Geometry, Signature, Biometric Database, Search

## 1 INTRODUCTION

In an increasingly digital world, reliable personal authentication is an important human computer interface activity. Biometrics such as fingerprints, face and voice verification are gaining industrial, government and citizen acceptance. The US-VISIT program uses biometric systems to enforce homeland and border security. Governments around the world are adopting biometric authentication to implement National ID and voter registration schemes [1]. FBI maintains national criminal and civilian biometric databases for law enforcement. In spite of the rapid proliferation of large-scale databases, the research community has thus far focused only on accuracy within small databases while neglecting the scalability and speed issues important to large-scale applications. In such applications, response time, search and retrieval efficiency also become important in addition to accuracy. Traditional databases index the records in an alphabetical or numeric order for efficient retrieval. In biometric templates, there is no natural order by which one can sort the biometric records.

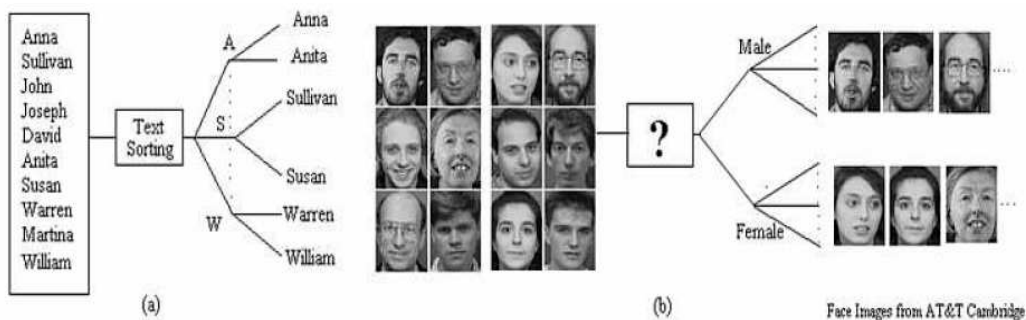


Figure 1: (a) Textual data indexed using traditional sorting methods, (b) Biometric data such as face images has no natural sequencing order and indexing and classification is non trivial.

We propose an original method for guiding the search in a biometric database, using the binning and pruning approach that perform a coarse level classification of the template before performing the exhaustive matching. The database is clustered / partitioned into several bins using the k-means clustering algorithm. During querying, the query template will be assigned to a certain partition of the database and all templates within this partition would be

considered as potential matches for the final level identification. To prevent the chances of missing the bin in which the actual match lies, we search  $C$  closest bins, where the value of  $C$  is closely related to the penetration rate of the system and is thus application dependent. Multiple biometrics can be used so that one biometric can be used to coarsely identify potential matches and a more reliable biometric is used to perform the final identification. In this paper we have used a parallel combination of hand geometry and signature biometrics for binning. The final search space is intersection of the templates identified as potential matches by both hand geometry and signature individually.

### 1.1 PROBLEM DEFINITION

It has been shown that the number of false positives in a biometric identification system grows geometrically with the size of the database [5]. If  $FRR$  and  $FAR$  indicate the false accept and reject rates during verification, then  $FRR_N$  and  $FAR_N$ , the rates of false rejects and accepts in the identification mode are given by

$$FAR_N = 1 - (1 - FAR)^N \quad (1)$$

$$\approx N \times FAR \quad (2)$$

$$FRR_N = FRR \quad (3)$$

$$\text{No. of false accepts} = N \times (FAR_N) \approx N^2 \times FAR \quad (4)$$

It can be seen that a biometric identification (1:N matching) system scales very poorly with the size of the database. A system that has acceptable performance during verification (1:1 matching) may become useless when used for identification across a large database due to this scaling behavior. There are two approaches in which we can attempt to reduce the error of such an identification system:

- (i) By reducing the  $FAR$  of the matching algorithm and
- (ii) By reducing the search space during identification

The  $FAR$  of a modality is limited by the recognition algorithm and cannot be reduced indefinitely if we are to accommodate any level of intra-user variance. Thus, a more practical approach to improve the speed and accuracy a biometric identification system is by reducing the number of records against which matching is performed during a query. This requires that the records in the database be classified, partitioned and indexed in some manner.

The effects of reducing the search space during identification are evident from the following mathematical analysis:

Assume that we are able to reduce the search space to a fraction  $P_{SYS}$  of the entire database. Then the resulting  $FAR$ ,  $FRR$  values and total number of false accepts is given by

$$FAR_{P_{SYS}N} = 1 - (1 - FAR)^{P_{SYS} \times N} \quad (5)$$

$$\approx (P_{SYS} \times N) \times FAR \quad (6)$$

$$FRR_{P_{SYS}N} = FRR \quad (7)$$

Figure 2 shows the effect that the reduction in search space has on the number of false accepts generated. It can be seen that when  $P_{SYS} \leq 0.1$ , the database scales very well with size. Thus, by restricting the final identification to be done with only a fraction of the entire database, we ensure not only good response time, but also improved accuracy of the overall system.

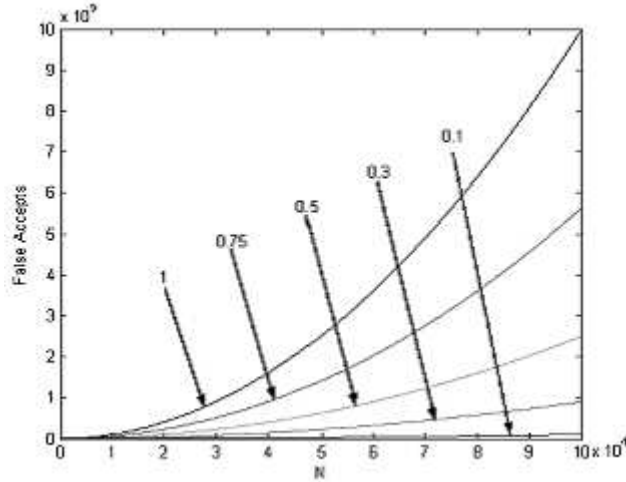


Figure 2: Effect of the penetration factor  $P_{SYS}$  on the number of false accepts

## 1.2 PRIOR RELATED WORK

There exist well-established procedures such as Henry classification system [3] to classify fingerprint records. In this system, a fingerprint record containing impression of 10 fingers is classified into 1024 separate classes or bins based on the type of ridge pattern such as 'whorl', 'loop', 'arches', 'tented arch' etc. on each finger. Classifying Fingerprints into the Henry classes has been tried by Jain et al in [7], yielding a system with 12.4% FRR. A similar work by Ratha et al in [8] yielded a FRR of 10% with search space pruned to 25% of the original database. However, in an experiment conducted by Cappelli et al [9] on NIST Special Database – 4, it was shown that the distribution of fingerprint population was not uniform, with 2 of the 5 Henry classes they considered holding nearly 65% of the population. Besides, such natural methods for classifying the biometric samples do not exist for other biometric modalities such as hand geometry and signature. Furthermore, even Henry classification scheme cannot be applied to individual prints. Thus a more sophisticated generalized methodology for classifying the biometric templates is needed for achieving efficient search and retrieval from the biometric databases.

## 2 BINNING AND PRUNING

We propose to reduce the search space by partitioning the database into several bins. Following such binning, the biometric database will be partitioned such that the templates in each bin are similar and correspond to some natural or statistical class. In general, a biometric template  $X_i$  can be represented as a  $k$  dimensional vector  $[x_1, x_2, \dots, x_k]$ . The problem of binning is essentially that of classifying the  $N$  vector templates in the database constituting a vector space  $S$  into  $M$  distinct classes  $\{Y_1, Y_2, \dots, Y_M\}$ , such that

$$\bigcup_{i=1}^N Y_i = S \text{ and } Y_i \cap Y_j = \Phi, \forall i \neq j \quad (8)$$

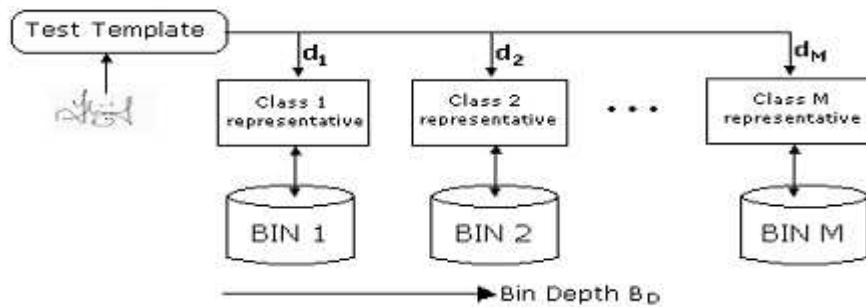


Figure 3: The test template would be tested with only the class representatives to identify the closest bin. Templates within the closest bin form the set of probable match

In case of the traditional 1:N comparisons for identification, the time needed for the system would be to determine the distance between the test template and the N templates in database. Thus the total time needed in such a case could be given as:  $\Theta(N)$ .

Using the binning approach, when a test template is to be identified, we have to simply find the C closest bins in which a probable match for the test template could be. To achieve this, we initially have to find the distance of the test template from all the bin centers. Furthermore, obtaining the C closest bin centers could be achieved in a single scan, needing at most  $C \times M$  number of comparisons. Thus the time complexity for determining the C closest bins could be given by  $\Theta(C \times M)$ . The total time for the identification task would hence be given as:

$$\Theta(C \times M) + C \times \Theta(A_M) \tag{9}$$

where  $A_M = \frac{N}{M}$  is the average bin density.

Now,  $C < M \ll N$  and  $A_M < N$ , thus,  $\Theta(C \times M) + C \times \Theta(A_M) < \Theta(N)$ , hence proving that the time needed for a binning system would be much lesser than that in a normal 1:N identification system. Further, the above analysis was done assuming a uniform distribution for all the bins. Even in the usual case of skewed distribution within the bins, the factor  $C \times \Theta(A_M)$  could be replaced by  $\Theta(P_{SYS} \times N)$ , in which case too our analysis of the time requirement would hold true since  $P_{SYS} \times N < N$ .

It is to be noted that vector representation may not be applicable to temporal biometric templates such as those found in speech. In such situations, Fourier Transform or discrete cosine transform (DCT) may be better used to obtain a vector representation.

## 2.1 MULTIMODAL BINNING

In this approach, we use multiple biometric to reduce the database search space. Figure 3 illustrates the proposed approach using hand geometry, signature and fingerprint modalities. Both hand geometry and signature templates are binned individually. During identification, the test template is used to identify the closest bins to which the user may belong. An intersection of templates within the bins of signature and hand-geometry form the final search space. A 1:N matching is then performed on this final search space to establish the identity of the individual. The binning may be accomplished by any space partitioning method. We use k-means clustering algorithm for binning the biometric modalities in our approach. The binning process divides the database D into partitions  $\{P_i\}$  such that if a template T belongs to partition  $\{P_i\}$ , then using a distance metric  $D(x,y)$  for the templates, we have

$$D(T, C_i) \leq D(T, C_j), \forall i \neq j \tag{10}$$

where,  $C_i$  represents the cluster centers of bins or partitions  $P_i$

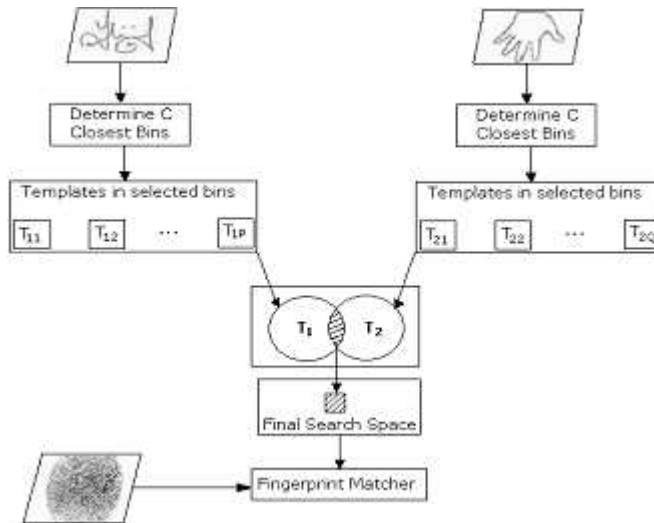


Figure 4: Schema for binning using multiple biometrics

For search space reduction we must have  $|P_i| \ll |D|$ . That is, the number of records in each bin has to be significantly less than the total number of records in the database. However as it will be seen in the section 2.1.1, this involves a trade-off between search space reductions and errors caused by bin miss. There are several other approaches in which multiple biometrics can be combined. In case of multiple biometrics the effective penetration rate is approximately given by  $P_{SYS} = P_{SYS1} \times P_{SYS2} \dots$  (11)

### 2.1.1 ACCURACY

Binning of the feature vectors is performed using the k-means clustering algorithm. Using the clustering algorithm, the set of input vectors  $[x_1, x_2, \dots, x_k]$  is divided into  $N$  bins  $\{Y_1, Y_2, \dots, Y_N\}$ . The k-means algorithm performs a hard-partition of the entire data-space. Thus the bins formed are such that:

Using this technique, the given template is quantized into a class  $Y_i$ . It is then compared with only those records that are also quantized into the same class.

Although binning reduces the penetration rate and consequently the false acceptance rate, it introduces another potential source of error. At the time of matching, if the bin in which the user record exists is missed, then a false reject is generated irrespective of the matching accuracy. If we define the accuracy of correctly binning the record as  $P(B)$ , the modified false accept and false reject are given by

$$FARN = FAR_{P_{SYS}N} \tag{12}$$

$$FRRN = (1 - P(B)) + P(B)FRR_{P_{SYS}N} \tag{13}$$

Therefore the process of binning increases the probability of false rejects in an identification system. Consider a screening application, where each international passenger compared against a watch-list of say terrorist suspects, a false reject means that a potential terrorist has been allowed to enter the country. One of the ways to reduce the percentage of false rejects is by including  $K$  nearest bins where the user's record is likely to lie instead of choosing the closest bin alone. This increases the  $P(B)$  at the cost of increasing the penetration rate  $P_{SYS}$ . Thus there exists a trade-off between the false reject rate and the fraction of database that has to be searched. The lower the false reject rate, higher is the search space we consider for the final matching. The relation between the two is evident in figure 9.

## 2.2 SIGNATURE FEATURES

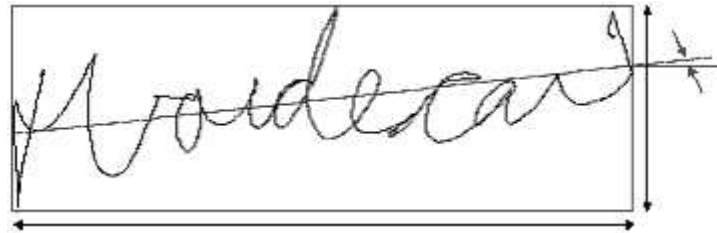


Figure 5: Height, width, major stroke length and angle features

The features selected from the biometric for binning should be such that they can be quickly and easily extracted to avoid any overhead due to binning. We captured the online signatures using a tablet. The data collected from an online signature usually includes the x-y coordinates and pressure. The signature data is thus a time series data. However, for the binning purposes we need a fixed template. Thus several features are extracted from the time-dependent signature data, which gives us the fixed-template representation for the signature biometric. The following features are selected - regression constants, compactness, signature length, major stroke length and angle, connected component count, hole count, stroke count, hole area and signing time [4].

Regression Constants: Signature data can be considered as a plot of x-y values. A simple linear regression can be done on this data to obtain the regression constants  $b_0$  and  $b_1$ , which are y-intercept and slope of the resultant line and can be evaluated as follows:

$$b_1 = \frac{\sum(x \times y) - n \times x_m \times y_m}{\sum x^2 - n \times x_m^2} \quad b_0 = y_m - b_1 \times x_m \quad (14)$$

**Compactness:** It is a measure of how elongated or compact the signature is. It is simply the ratio of the height to width of the bounding box for the signature.

**Signature Length:** Length is the cumulative distance between successive points in the signature. The pen up and pen down status should be used while calculating this distance. Length does not involve successive points with a different pen position status.

**Major Stroke Length and Angle:** Since signature is sampled at equal time intervals, the maximum velocity with which a person signs is captured by the major stroke length. This feature is usually non-varying for a person and very easy to extract. Once the length is found, the angle that the line makes with the horizontal can also be found adding more distinctiveness to the feature set.

**Connected Components Count:** The number of connected components in a signature varies from person to person. A point is said to belong to a connected component if it can be reached from any other point within that connected component. The number of connected components is found by pixel labeling method where points with same label or equivalent label belong to same component.

**Hole count and area:** Holes are enclosed spaces within a signature. If the pixels within a signature image are negated and the number of connected components is found, the end result is the number of holes in that signature. The hole area is found by counting all pixels with similar labels and adding all of them except the major hole that corresponds to the area enclosing a signature.

**Stroke count:** A stroke is considered as the part of a signature that is written without lifting the pen. In an online signature, the number of strokes is found by using the pen-up and pen-down status information provided by the tablet. Every pen-up constitutes the completion of a stroke.

**Signing time:** The time required to sign a signature is equivalent to number of points sampled by the tablet or the e-pad, provided that sampling is done at equal intervals.

### 2.3 HAND GEOMETRY FEATURES

The hand image is obtained by placing the palm on a Back-lit Translucent base. The light source beneath the translucent base helps in clearly distinguishing the palm from the background. Using a digital camera, we obtain the top-view of the palm. The Image is captured at a resolution of 640 X 480 dpi. The captured Image is then post-processed to obtain a binarized Image. From this binary Image we obtain the longest contour by using an algorithm similar to the chain code contour extraction method in [11]. Thus, the algorithm to extract the feature involves the following steps:

(1) Image acquisition, (2) Converting to binary image, (3) Contour extraction and noise elimination and (4) Feature extraction as shown below:

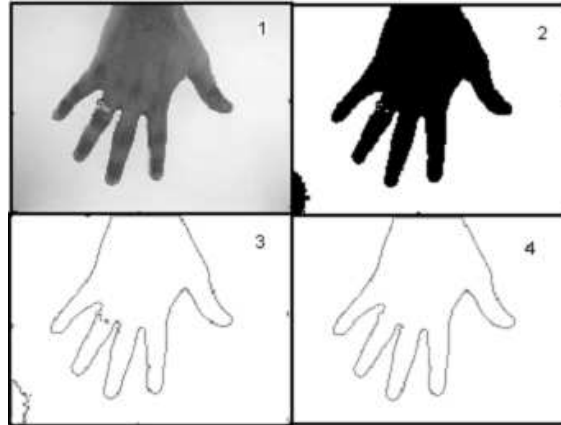


Figure 6: Showing the 4 stages for feature extraction in hand geometry

Being a Peg-less system, there are no fixed axes along which to measure the features, making feature extraction a challenging task. To determine the points defining the axes along which to measure the features we first find the tip and the valley of each finger along the curvature. The various features that we are interested in involve the lengths and widths of the various parts of the fingers, ratios of heights to widths of certain fingers. All together we have identified 27 invariant feature of the hand [6] shown below.

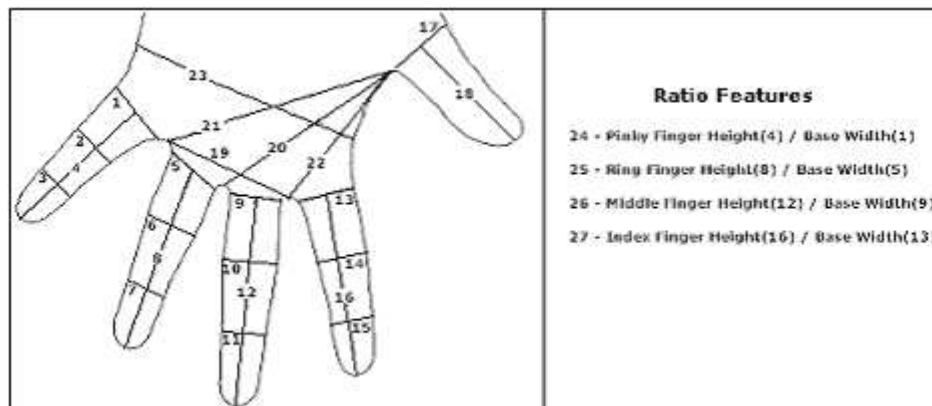


Figure 7: Showing the 27 hand geometry features used

#### Distance Features:

**Widths:** The base widths of all the 5 fingers are considered. The base width is width of the finger measured just before the start of the webbing of each finger. The remaining width features are measured for all fingers except the thumb. These remaining width features are measured at approximately the 2 points of inflection in each finger. Thus in all, there are 8 inflection widths, 2 per finger, constituting a total of 13 width features.

**Heights:** The height of a finger is the distance measured from the center of the base of the finger to the tip-point of the finger. The height is measured for all the fingers.

The remaining distance features consist of the palm width and the distances between various valley-points found at the webbing between fingers as shown in the above figure.

#### Ratio Features:

The ratios of finger heights to widths of all fingers except the thumb are considered. The ratio features allow us a little tolerance in locating the valley-points of the fingers, since if we measure the base-width of the finger a little lower than the actual location, the base width would be lesser than the actual base width of the finger. Also, the height of the

finger would then be correspondingly lower, thereby making the ratio of the height to width nearly constant. Similar argument holds when we measure the base-width a little farther down, thereby increasing the measured base-width and hence also the corresponding height of the finger.

### 2.4 PERFORMANCE EVALUATION

The hand geometry and the signature features are normalized by using Z-Score [10] method, which takes into account mean and standard deviation for minimizing the effects of outliers. Once the features are normalized, clustering is then performed on them and bins are formed. Each bin is represented only by its bin center and distance is measured with respect to this center alone. We thus derived different number of representative vectors for both modalities through the k-means clustering algorithm. We used signature and hand-geometry templates of 50 users. We used 5 templates per user for training and 5 separate templates for testing. We then evaluated the algorithm based on the percentage of the database ( $P_{SYS}$ ) that has to be searched in order to find the given user's record. We select not only the closest bin but also its two closest neighbors during the search process. This procedure ensures that we have 0% ( $P(B)=1$ ) FRR or no bin misses all the time. The effectiveness of binning can be demonstrated by using a validation set to determine the bin penetration rate, which is described as

$$P_{SYS} = \frac{N_S \times T_B + N_B}{N_D} \tag{15}$$

Here,

$N_S$ : the number of bins to be searched such that  $P(B)=1$ ,       $N_B$ : the number of bins,  
 $T_B$  is the average templates per bin                                       $N_D$  is the database size.

On performing a parallel combination of the binning of the 2 individual biometrics and assigning the final search space as the intersection of the set of probable matches determined by the 2 binning biometrics, we obtained a penetration rate as low as 5%. The penetration for the combination method has been found by simply taking the intersection of templates in the set of potential matches for the 2 biometric individually. The effect of varying the number of bins using the parallel combination on the penetration rate is shown below.

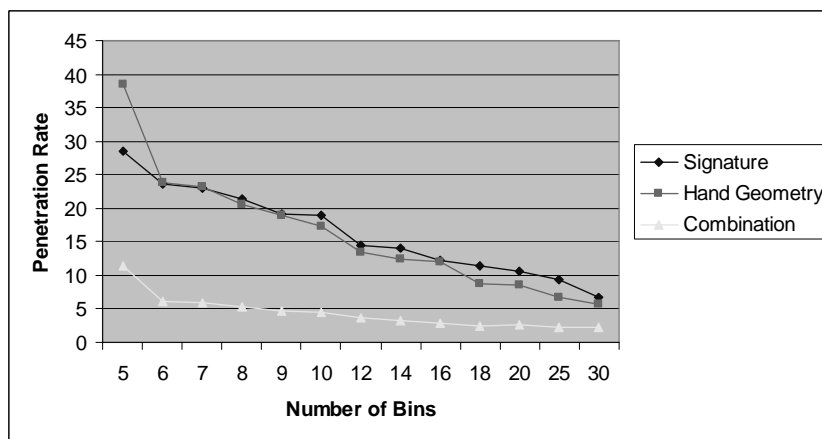


Figure 8: Binning results for the combination method

### 3 CONCLUSION AND FUTURE WORK

We have presented a framework for binning in biometric database. Binning was achieved using the k-means clustering algorithm. Both hand geometry and signature biometric data were used for the binning purposes. The experiments were run, initially using both the biometric individually and then using a parallel combination of the two biometric. We experimentally showed that multiple biometrics yields a great reduction in search space than using a single biometric, without affecting the FRR during identification.

Binning the database using the k-means clustering algorithm needs specifying the number of clusters/bins that the data would be partitioned into. As the data size increases, the task of clustering the entire database would take a long time. If the database is a static one, then binning the entire database is simply a one-time process and would be quite affordable. For a real-time system, as newer templates would be introduced into the system, we would have to consider these templates too into our search space. However, it is not feasible to re-cluster the entire database with the incoming of every single template. Thus for a real-time system, we currently propose to have a separate 'default' bin, which would store all the new incoming templates. Thus along with the closest bins, the 'default' bin too would have to be searched for the identification task. The new templates would be stored in the 'default' bin as long as the size of the bin is below a set threshold size else the entire database would be re-clustered. This prevents us clustering the data for every new template. Our future work would look at employing methods of online-clustering for the binning approach.

Our future work would also involve generation of synthetic data and evaluation of the proposed scheme over very large biometric databases.

### **ACKNOWLEDGEMENT**

We thank Pavan Rudravaram for his role in data collection and implementation of the contour extraction code.

### **REFERENCES**

1. Ruud Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. Guide to Biometrics, Springer Verlag 2003
2. G. M. Gray. Vector quantization, IEEE ASSP Magazine, pages 4–29, 1984.
3. International Biometric Group. The henry classification system. <http://www.biometricgroup.com/>.
4. Srinivas Palla. Search space reduction by binning and pruning in context of biometrics. Master's thesis, Department of Computer Science and Engineering, University at Buffalo, 2004.
5. D. Maio, D. Maltoni, A. K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer Verlag, 2003.
6. Pavan Rudravaram. Prototype pegless hand geometry verification. Technical report, Center for Unified Biometrics and Sensors, University at Buffalo, 2004.
7. Anil Jain, Sharath Pankanti. Fingerprint Classification and Matching, Handbook for Image and Video Processing, A. Bovik (ed.), Academic Press, April 2000
8. Ratha N., Karu K., Chen S., Jain A.. A Real-time Matching System for Large Fingerprint Databases. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 8, August 1996.
9. Cappelli R., Maio D., Maltoni D., Nanni L. A two-stage fingerprint classification system. ACM SIGMM workshop on Biometrics methods and applications, 2003
10. Jiawei Han, Micheline Kamber, "Data mining concepts and techniques", Morgan Kauffmann (2001).
11. V. Govindaraju, Z. Shi and J. Schneider, "Feature Extraction Using a Chain coded Representation of Fingerprint Images", International conference on Audio- and video-based biometric person authentication (AVBPA), surrey, UK, 2003