International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 30927-30942

http://www.ripublication.com

Clustering of Ears based on Similarity Metrics for Personal Identification

M.A.Jayaram[1], Prashanth.G.K.[2], Divya K.M[3]

[1]Director , [2]Assistant Professor ,[3]Post Graduate Student
Department of MCA,
Siddaganga Institute of Technology, Tumkur, Karnataka, India
jayaram_mca@gmail.com
prashanthgk@gmail.com
divyakm51@gmail.com

Abstract

Ear biometrics has been found to be a good and reliable technique for human recognition. Due to significant advantages, ear biometric has gained momentum. In this direction, we propose a method to cluster ears based on similarity measures and to make use of this for quick retrieval of the image followed by the personal identification. This work involves elicitation of shape based features like distribution of planner area, moment of inertia with respect to minor and major axis and radius of gyration with respect to minor and major axis from ear images. We use four similarity measures for clustering 605 ear images. The method involves two phases,

i. Determining centroids of predetermined groups by k-means clustering, and
ii. Using so obtained centroids to refine the clusters using similarity measures.

From the computational experiments carried out on 605 ear images, it is revealed that Cosine, Dice and Jaccard similarity were able to effectively cluster the image database into three groups. However, Overlapping similarity measures ended up only in two groups. The cluster analysis showed comparatively high values of entropy, purity, specificity, precision, recall and F-measure respectively for Jaccard, Dice and Cosine similarity function. The image retrieval rate became faster by by an average of 12.33% when database was organised in cluster groups when compared with image retrieval time with unorganised data followed by recognition accuracy of 92.5%.

Keywords – kmeans clustering, Jaccard, Cosine, Dice, Overlapping, shape based Ear features, Moment of Inertia.

I. INTRODUCTION

Personal identification is receiving renewed importance day-in and day-out with evolving security systems. Most of the traditional identification methods, which are widespread in the commercial systems, have many disadvantages. For example, personal identification number (PAN), typing logins, passwords, displaying identification cards or using specific keys require users to take active part in the process of identification. Above all, traditional methods are unreliable because it is hard to remember PINs and passwords, and it is fairly easy to lose ID cards and keys. Biometric methods easily deal with these problems because users are identified by who they are, not by something they have to remember or to carry with them[1].

Human ears are used as main features in the forensic science from many years. It is reported that ear prints found on crime scene have been used as proof in hundreds of cases in US and Netherlands[2].

Human ears have many distinguish features. It doesn't change considerably during human life, there will be no change in the ear in its configuration even if a person is in emotion. Ear features are relatively fixed and invariant [3].

The ear biometric based person identification as gained increased interest among the researchers. This is because the detailed structure of the ear is not only unique from person to person but also permanent. The appearance of the ear doesn't change over the course of human life. Added to this the acquisition of ear images doesn't necessarily require but nevertheless considered to be non-inclusive [4].

Similarity measures are the functions, which compute the degree of similarity between the pair of objects. There is large number of similarity coefficients found in the literature probably because the best similarity measure doesn't exist yet. In this work, we elaborate application of widely used similarity functions like Cosine, Jaccard, Dice and Overlapping similarity. The case in point is clustering the ear images based on similarity measures and finally making personal identification system.

The rest of the paper is organized as follows. Section II discusses some of the related works. Section III briefs on the novel shape based biometric. Section IV elaborates about the data used in the model. A brief presentation about four similarity metrics is in section V. The details of computational experiments is explained in section VI. Analysis of the results is done in section VII. Section VIII concludes the paper.

II. RELATED WORKS

A new method based on a combination of supervised and unsupervised learning for clustering data without any preliminary assumption on the cluster shape is implemented for Iris dataset. This is obtained by extracting the dissimilarity relations directly from the available data [5].

A novel approach directed towards the automatic clustering of x-ray images has been attempted. The clustering was carried out based on multi-level feature of given x-ray images such as global level, local level and pixel level. The approach involves a combination of k-means and hierarchical clustering techniques this work has reported for having shown high level of accuracy [6].

Xi Cheng et al [7] have used similarity measures in multi-sample biometric systems. Both Pearson's correlation and Cosine similarity are used. Computational experiments have shown a better performance than using raw matching scores.

Roman V. et al [8] have compared performance of similarity measure functions to that obtained from customized field-specific approach in the domain of strategy-based behavioral biometrics. While all similarity measure functions showed a relatively high accuracy levels during user verification, weighted Euclidian similarity measures has slightly outperformed than general approaches such as Manhattan distance or Mahalanobis distance as claimed.

Satya Chaitanya Sripada et al [9] have compared the for K-means and Fuzzy C means clustering using the Purity and Entropy. The paper reported that, The K-means has lower value of purity and high value of entropy compared to Fuzzy C Means. The Fuzzy C means clustering is more accommodating for medical data sets when compared to K means.

Vikas Thada et al [10] have focused on comparative analysis for finding out the most relevant document for the given set of keywords by using three similarity measures viz Jaccard, Dice and Cosine similarity measures by using genetic algorithm approach. Due to the randomized nature of genetic algorithm the best fitness value is the average of 10 runs of the same code for a fixed number of iterations. The result states that the best fitness values were obtained using the Cosine similarity coefficients followed by Dice and Jaccard.

III. SHAPE BASED BIOMETRICS

In this work, a novel idea that makes use of planner surface area properties has been used. For this, a ear is considered to be a planner surface. The moment of inertia(MI) and its related five parameters are elicited from the ear images. The features considered are given in Table 1. The details of the features, their extraction, the evaluation and the development of the system for human identification making use of these features is elaborated in the seminal work of authors [11]. However , for the sake of completeness the features are briefly explained in the following paragraphs.

The surface area of the ear is the projected area of the curved surface on a vertical plane. Moment of Inertia (MI) is the property of a planar surface which originates whenever one has to compute the moment of distributed load that varies linearly from the moment axis. Moment of Inertia is also viewed as a physical measure that signifies the shape of a planar surface and it is proved that by configuring the shape of planar surface and hence by altering the moment of inertia, the resistance of the planar surface against rotation with respect to a particular axis could be modulated or altered [12]. Therefore in this work, moment of inertia of ear surface with respect to two axes i.e. the major axis and the minor axis are considered to be the best biometric attributes that could capture the shape of irregular surface of the ear in a more scientific way.

Further, major axis is the one which has the longest distance between the two points on the edge of the ear, the distance here is the maximum among point to point

Euclidean distance. The minor axis is drawn in such way that it passes through tragus and is orthogonal to the major axis. Therefore, with different orientation of ears the orientation of major axis also changes. Being perpendicular to major axis, the orientation of minor axis is fixed.

The projected area is assumed to be formed out of segments. The area of an ear to the right side of the major axis is considered to be made out of six segments. Each of the segments thus subtends 30^0 with respect to the point of the intersection of the major axis and minor axis. The extreme edge of a sector is assumed to be a circular arc. Typical ear edge with measurements is shown in Figure 1.

The measurements are

- θ → Inclination of the central radial axis of the segment with respect to minor axis (in degrees).
- r → The length of the radial axis (in mm).

The conversion of number of pixel into linear dimension (in mm) was based on the resolution of the camera expressed in PPI (Pixel Per Inch). In this work 16Mega pixel camera, at 300 PPI was used. The computation of linear distance is straight farward mm=(number of pixel*25.4)/PPI [1 inch=25.4 mm]. With these measurements, the following parameters are computed.

Moment of inertia with respect to minor axis I_{min}

$$I_{min} = \sum_{i=1}^{6} a_i y_i^2 \qquad (1)$$

Where a_i is the area of a the i^{th} segment and y_i is the perpendicular distance of the centroid of the i^{th} segment with respect to minor axis.

$$a_i = \theta r^2 \qquad (2)$$

$$y_i = C \sin\theta \qquad (3)$$

Here, C is the centroidal distance of the segment with respect to the intersection point of the axes, which is given by[13];

$$C = \frac{2}{3}\frac{r \sin\theta}{\theta} \qquad (4)$$

Simlarly, moment of inertia with respect to major axis I_{max} ,x_i is the perpendicular distance of the centroid of the ith segment with respect to major axis.

$$I_{max} = \sum_{i=1}^{6} a_i x_i^2 \qquad (5)$$

Where $x_i = C\cos\theta$ (6)

From the computed values of moment of inertia and area of the ear surface, the radii of gyration with respect to minor axis(RGx)and major axis(RGy) were computed. The formulae for radii of gyration are given by[14].

$$RGx = \sqrt{\frac{I_{min}}{A}} \tag{7}$$

$$RGy = \sqrt{\frac{I_{max}}{A}} \tag{8}$$

Where , A is the sum of areas of six segments.

$$A = \sum_{i=1}^{6} a_i \tag{9}$$

Radius of gyration is the distance from an axis at which the mass of a body may be assumed to be concentrated and at which the moment of inertia will be equal to the moment of inertia of the actual mass about the axis. It is also equal to the square root of the quotient of the moment of inertia and the mass.

Table 1: Ear Shape Based Features in Classification

Sl. No	Attributes
1	Area (mm^2)
2	Moment of Inertia Y (Imax) (mm^4)
3	Radius of gyration Y (RGy) (mm)
4	Moment of Inertia X (Imin) (mm^4)
5	Radius of gyration X (RGx) (mm)

IV. DATA FOR THE MODEL

Ear images for this work were acquired from the pupils of Siddaganga group of institutes. The subjects involved were mostly students and faculty numbering 605. In each acquisition session, the subject sat approximately one meter away with the side of the face in front of the camera in outside environment without flash.

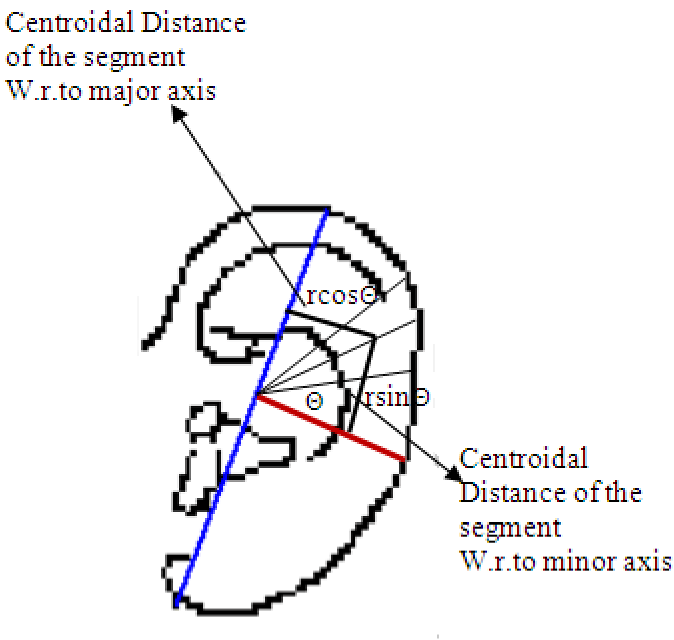
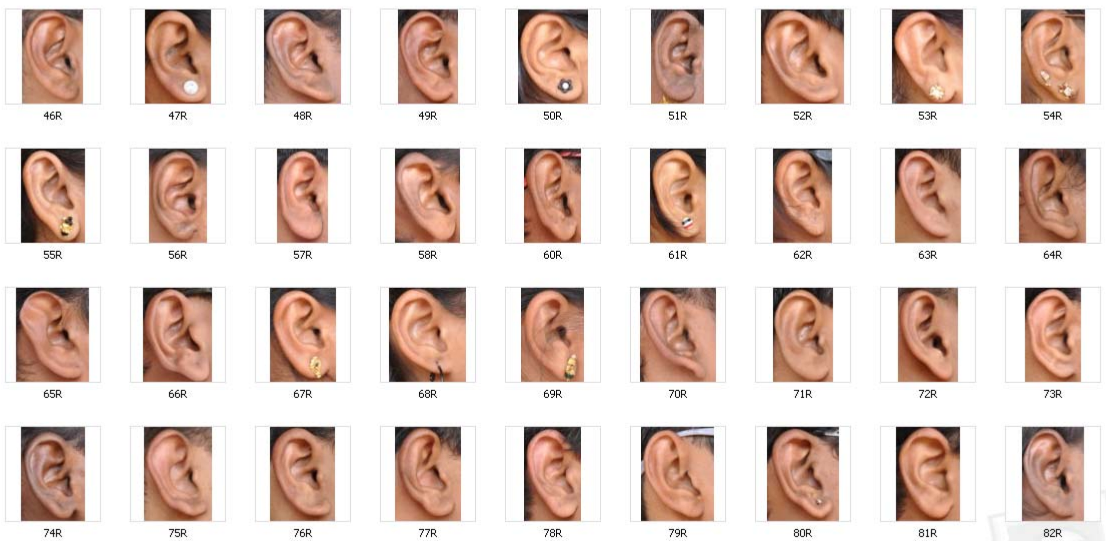


Fig. 1. Typical ear edge with M.I. parameters.

The images so obtained were resized in such a way that only ear portion covers the entire frame having pixel matrix.

A segment of the database involving right ears of persons is shown in Fig 2.



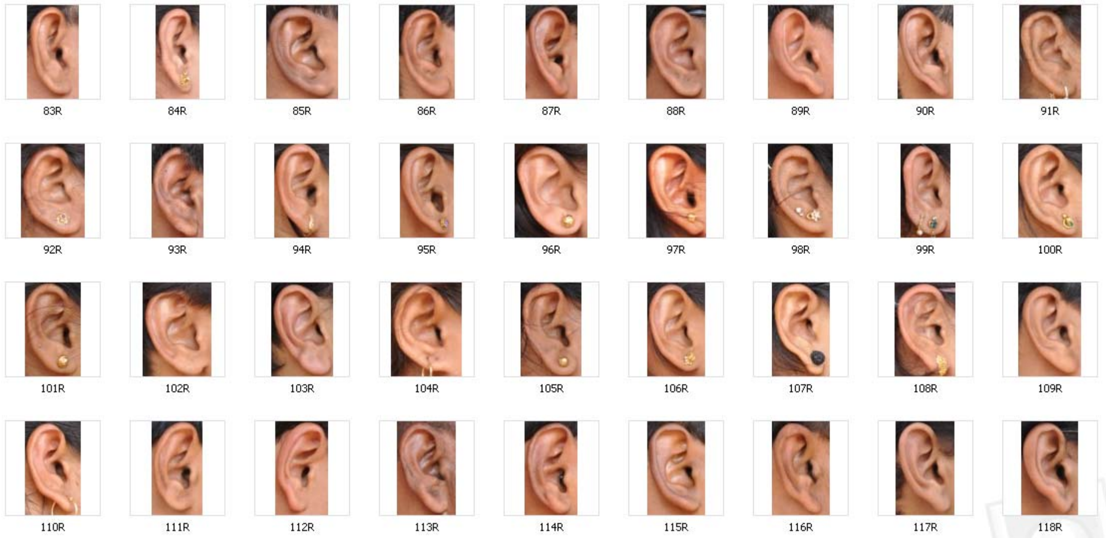


Fig.2. A sample gallery of right ear database

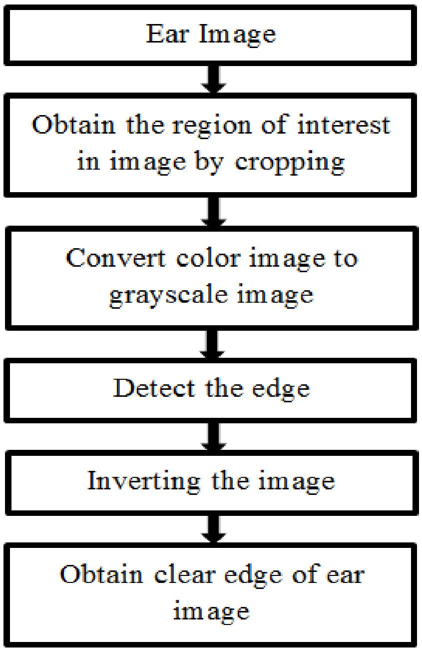


Fig 3. The Steps involved in ear edge extraction.

The color images were converted into gray scale images followed by uniform distribution of brightness through histogram equalization technique. The delineation of outer edge of each ear was obtained using canny edge detection algorithm. The resulting edge was inverted to get a clear boundary shape of the ear. The conceptual presentation of the process involved is shown in Figure 3.

V. SIMILARITY MEASURES

Before clustering, a similarity/distance measure must be determined. The measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. In many cases, these characteristics are dependent on the data or the problem context at hand, and there is no measure that is universally best for all kinds of clustering problems [15]. Moreover, choosing an appropriate similarity measure is also crucial for cluster analysis, especially for a particular type of clustering algorithms. For example, the K-means clustering algorithms [16].

In general, similarity/distance measures map the distance or similarity between the symbolic descriptions of two objects into a single numeric value, which depends on two factors— the properties of the two objects and the measure itself. Different measure not only results in different final partitions, but also imposes different requirements for the same clustering algorithm [17].

A. Metric

Similarity measures are essential to solve many pattern recognition problems such as classification, clustering, and retrieval problems. Similarity between two data points (feature vectors or the rows of the data matrix) is measured through a function $f_s(x,y)$ which is also called *proximity measure* [18].

A function $f_s(x,y)$ is required to satisfy following identities for $x,y \varepsilon R^p$

i. $f_s(x,y) = f_s(y,x)$ (10)

ii. $f_s(x,y) \leq f_s(x,x)$ *(11)*

iii. $f_s(x,z) \leq f_s(x,y) + f_s(y,z)$ *(12)*

iv. $f_s(x,y) \geq 0$ *(13)*

A function $f_s(\,)$ is called a normalized similarity measure if:

$$f_s(x,x) = 1 \qquad (14)$$

When two binary feature vectors are involved, the similarity would exist if many 1's coincide. This observation may be represented by a product , so that the scalar product of feature vectors will be reasonable. For real valued feature vectors, the similarity measures are based on scalar products that may be normalized in different ways. A few of them are presented.

B. Cosine Similarity

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the Cosine of the angle between them [19]. The Cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

$$f(x,y)=\frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} {x_i}^2 \sum_{i=1}^{p} {y_i}^2}} \tag{15}$$

One of the reasons for the popularity of Cosine similarity is that it is very efficient to evaluate, especially for sparse vectors, as only the non-zero dimensions need to be considered.

C. Jaccard Similarity

The Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of sample sets [20]. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$f(x,y)=\frac{\sum_{i=1}^{p} x_i y_i}{(\sum_{i=1}^{p} {x_i}^2+\sum_{i=1}^{p} {y_i}^2-\sum_{i=1}^{p} x_i\, y_i)} \tag{16}$$

D. Overlapping Similarity

The overlapping coefficient (or, Szymkiewicz-Simpson coefficient) is a similarity measure related to the Jaccard index that measures the overlap between two sets, and is defined as the size of the intersection divided by the smaller of the size of the two sets [21]:

$$f(x,y)=\frac{\sum_{i=1}^{p} x_i y_i}{Min(\sum_{i=1}^{p} {x_i}^2, \sum_{i=1}^{p} {y_i}^2)} \tag{17}$$

If set *X* is a subset of *Y* or the converse then the overlap coefficient is equal to one.

E. Dice Similarity

The Dice Similarity coefficient of two vectors is twice the sum of dot product of the vector divided by the sum of the second degrees of the vectors [22]. It is given by:

$$f(x,y)=2\,\frac{\sum_{i=1}^{p} x_i y_i}{(\sum_{i=1}^{p} {x_i}^2+\sum_{i=1}^{p} {y_i}^2)} \tag{18}$$

VI. METHODOLOGY

Computational Experiments were carried out to find the appropriate number of clusters using K-Means. The algorithm settled for three distinct clusters with a minimum overlapping. The distribution of the data in three clusters by K-Means shown in Figure 4.

The centroids of these clusters were used for measuring similarity with all the ear images in the database. This comparison was done using four similarity measures to obtain refined grouping. A conceptual presentation of the methodology used depicted through a block diagram shown in Figure 5.

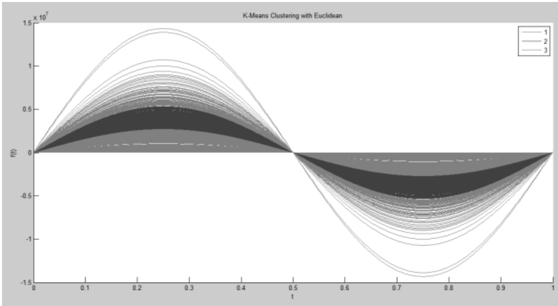


Fig 4: Distribution of data in 3 groups by K-means.

It is very difficult to conduct a systematic study comparing the impact of similarity metrics on cluster quality, because objectively evaluating cluster quality is difficult in itself. In practice, manually assigned category labels are usually used as baseline criteria for evaluating clusters. As a result, the clusters, which are generated in an unsupervised way, are compared to the pre-defined category structure, which is normally created by human experts. This kind of evaluation assumes that the objective of clustering is to replicate human thinking, so a clustering solution is good if the clusters are consistent with the manually created categories. However, in practice datasets often come without any manually created categories and this is the exact point where clustering can help [23].

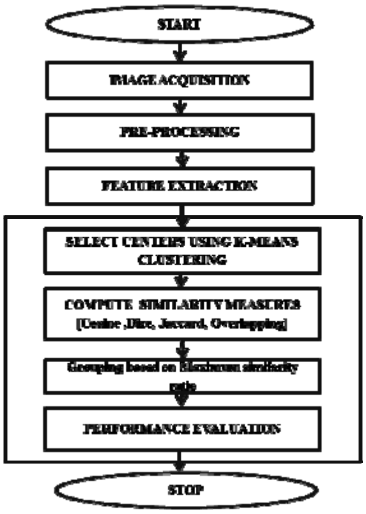


Fig 5: Block diagram of the model.

A. DATASET

The 605 data sets with five features are chosen for the computational experiments. The 2-norm of a vector is used to normalize the entire dataset. A sample segment of the dataset with features is shown in Table 2.

Table 2: The sample dataset

Sl. No	Area	Imax	RGy	Imin	RGx
1	131.3487	195355.9	38.56563	50.89822	0.622499
2	135.0409	371368.5	52.44088	140.7316	1.020853
3	404.5635	3158077	88.35233	766.4084	1.376375
4	241.7773	744226.9	55.48108	116.0455	0.692798
5	370.4741	2691695	85.23815	310.3129	0.91521
6	272.2138	3054815	105.9345	0.344449	0.035572
7	358.0337	3395618	97.38621	254.3965	0.842934
8	369.2937	2464924	81.69882	264.376	0.846107
9	217.2377	2858884	114.7178	11.01563	0.225184
10	360.2648	2844322	88.8543	641.5627	1.334469
11	338.5039	1991710	76.70634	368.703	1.043654
12	379.9424	5368573	118.8695	240.5883	0.795753
13	412.1489	4025855	98.833	316.661	0.876537
14	639.5815	9003344	118.6462	3243.304	2.251883
15	376.3808	2785671	86.03025	647.0868	1.311196
16	435.8933	4651435	103.3007	822.7214	1.37384
17	369.8076	2818062	87.2946	258.6729	0.836349
18	266.0732	1406845	72.71474	437.977	1.282995
19	441.7652	7186791	127.5474	537.809	1.103363
20	450.2947	4146650	95.96222	1295.909	1.696441
21	405.7415	3655038	94.91202	237.0266	0.764318
22	414.6593	4291438	101.7316	1138.657	1.657108
23	439.04	5612504	113.0645	1007.022	1.514494
24	569.4626	7267216	112.967	1253.937	1.483902
25	388.5733	4062368	102.2476	810.9251	1.444621

B. EVALUTAION

Each of the 605 data instances were ran through K-Means algorithm and centroids for three groups were obtained. This is sown in Table 3. This was the starting point for the application of the similarity functions. The so obtained centroids were compared against each of the feature data to find the similarity measures. Further, regrouping of the database was made using the similarity measures obtained. This process of re-clustering resulted in assorting of the database into three groups. The distribution of the database as evinced by the four methods is presented in table 4.

We justify the effectiveness of proposed similarity measures by using standard cluster quality metrics like Purity and Entropy. The greater the value of purity indicates good clustering. The entropy is a negative measure, the lower the entropy the better clustering it is [9]. Measures must always facilitate for the increase of purity. Entropy is more reliable in gauging the effectiveness of similarity measure as it considers the overall distribution of all categories in the clustering results. Further, the purity and entropy are independent of the actual results of the clusters. Even a pair of clusters produced by two different similarity measures can have purity measure very close and their entropy can be used to decide the effectiveness of the similarity function, if it has the lowest entropy value [9].

i. Purity : Purity can be defined as the maximal precision value for each class j, The purity for a cluster j can be computed as:

$$Purity(j) = \frac{1}{c_j} max\,(c_{ij}) \tag{19}$$

Where c_j is total number of data objects belongs to cluster j.

We then define the purity of the entire clustering result as:

$$Purity = \sum_j \frac{c_j}{N} purity(c_j) \tag{20}$$

Where $\sum_j c_j$ i.e. the sum of the cardinalities of each cluster, Note that we use this quantity rather than the size of the document collection for computing the purity.

ii. Entropy: The entropy measure evaluates the distribution of categories in a given cluster. The entropy for a set of cluster is defined as:

$$H=-\sum i\ p_i \log p_i \tag{21}$$

Where p_i is probability of cluster i.

We need to maximize the purity measure and minimize the entropy of clusters in order to accomplish high quality clustering results.

The other evaluation measures are, the Precision, Recall, specificity and F-Measures. Precision measures the exactness of a classifier, whereas, recall measures the completeness, or sensitivity of a classifier. Improving recall often decreases precision and vice versa. Precision and recall in combination evolve into F-measure, which is the weighted harmonic average of precision and recall.

VII. EXPERIMENTAL RESULTS AND ANALYSIS

Table 5,6 and 7 describes the confusion matrix of Cosine , Jaccard, Dice and Overlapping Similarity function. Accuracy is the percentage of predictions that are correct. Precision gives the percentage of positive predictions that are correct. Recall says about the percentage of positive labelled instances that were predicted as positive.

Specificity is the percentage of negative labelled instances that were predicted as negative.

In the case of Jaccard similarity, precision is at 91%, recall: 93%, specificity: 89%, F-measure: 92% and accuracy: 92% these values suggest that this is quite a reasonable grouping with respect to all other similarity measures. The performance of Dice similarity is completely aligned with Jaccard similarity measure.

Table 3: Centroids of clusters as determined by K-means algorithm

	Area	Moment of Inertia X	Radius of gyration X	Moment of Inertia Y	Radius of gyration Y
Group I	270.4792	1666994	74.8501	292.8828	0.877
Group II	396.5784	3801852	98.4193	683.1179	1.156
Group III	501.6093	6866031	118.308	1605.915	1.5851

Table 4: k-means ,Cosine, Jaccard, Overlapping and Dice clusters.

	Group I	Group II	Group III
K-means	247	278	80
Jaccard	213	300	92
Cosine	177	230	198
Dice	213	300	92
Overlapping	265	0	340

In order to expeditiously make use of similarity based groups the test images were organized in three consecutive blocks involving 213, 300 and 92 ear images in first, second and third groups respectively. Test images were identified against the template images and CPU time was measured. For the sake of comparison, the test images were also matched in the database which was disorganized and CPU times were again measured. In this experiment, it was found that the average decrease in CPU time involved for image recognition and retrieval was around 12.33% with organized database. The details of which is shown in Table 5.

Table 5. Average CPU time

No. of test images	With unorganized database (secs)	With organized database (secs)
200	0.0733	0.0631

Table 5: Confusion matrix of Cosine Similarity

	C1	C2	C3
C1	167	49	31
C2	0	49	31
C3	10	100	168

Table 6: Confusion matrix of Jaccard Similarity

	C1	C2	C3
C1	213	0	34
C2	0	80	0
C3	0	12	266

Table 7: Confusion matrix of Dice Similarity

	C1	C2	C3
C1	213	0	34
C2	0	80	0
C3	0	12	266

Table 8: Confusion matrix of Overlapping Similarity

	C1	C2	C3
C1	0	247	0
C2	80	0	0
C3	185	93	0

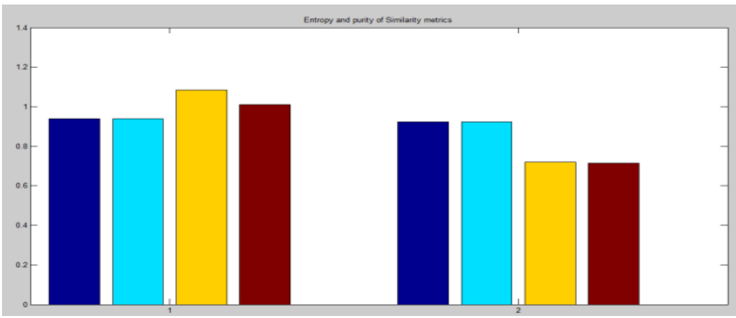


Fig. 6. Entropy and Purity of similarity measures

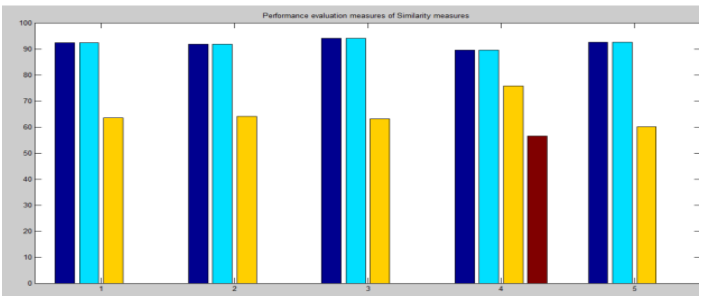


Fig. 7. Accuracy, Precision, Recall, specificity and F-Measures of Similarity measures

Table 9. Performance evaluation measures of Similarity measures

	Jaccard	Dice	Cosine	Overlapping
Entropy	0.9403	0.9403	1.0847	1.0093
Purity	0.9240	0.9240	0.7190	0.7140
Accuracy (%)	92.40	92.40	63.47	0
Precision (%)	91.87	91.87	64.00	0
Recall (%)	93.97	93.97	63.09	0
Specificity (%)	89.56	89.56	75.84	56.59
F-measure (%)	92.55	92.55	60.05	0

VIII. CONCLUSIONS

This paper presented application of similarity measures for obtaining the refined clustering of ear images with shape based features. Based on the research work, following conclusions are drawn:

- Similarity metrics could be effectively used for clustering biometric database in general, and ear biometric in particular.
- Cosine, Dice and Jaccard similarity measures were able to effectively cluster the image database into three groups with minimum overlapping as seen from the confusion matrix.
- The cluster analysis showed comparatively high values of purity, specificity, precision, recall and F-measure respectively for Jaccard, Dice and Cosine similarity functions. However, overlapping similarity function was inefficient and ended up in generating only two clusters.
- The decrease in the time of image matching and retrieval of personal details is significant when database is organized as per the groups when compared to identification time with unorganized database.

IX. REFERENCES

[1] Michal choras, "Ear Biometrics Based on Geometrical Feature Extraction". Electronic Letters on Computer Vision and Image Analysis 5(3):84-95, 2005.

[2] Hoogstrate A.J., Heuvel van den., Huyben E., " Ear Identification Based on Surveillance Camera's Images", Netherlands Forensic Institute,2000.

[3] Iannarelli, A., "*Ear Identification",* Forensic Identification Series, Paramount Publishing Company, Fremont, California (1989).

[4] Anika Pflug, Christoph Busch, "Ear Biometrics: A Survey of Detection, Feature Extraction and Recognition Methods".

[5] Satish R. Kolhe, Ranjana S. Zinjore, "Clustering Iris Data using Supervised and Unsupervised Learning" , International Journal of Computer Science and Application Issue 2010, ISSN 0974-0767.

[6] Chhanda Ray, Krishnendu Sasmal, "A New Approach for Clustering of X-ray Images" , IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8, July 2010 ISSN (Online): 1694-0784 ISSN (Print): 1694-0814

[7] Xi cheng, Sergey Tulyakov and Venu Govindaraju, “Utilization of Matching Score vector Similarity Measures in Biometric Systems” , 978-1-4673-1612-5/12/$31.00 2012 IEEE.

[8] Roman V, Yampolskiy and Venu Govindaraju, “Similarity Measure Function for Strategy-Based Biometrics”, World Academy of science,Engineering and Technology 24 2006.

[9] Satya Chaitanya Sripada Dr. M.Sreenivasa Rao , “Comparison of purity and entropy of k-means clustering and fuzzy c means clustering”, Indian Journal of Computer Science and Engineering (IJCSE).

[10] Vivek Jaglan, Vikas Thada , “Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm”, published in International Journal of Innovations in Engineering and Technology (IJIET)

[11] M.A.Jayaram, Prashanth.G.K., Sachin.C.Patil,(2015) “Inertia based ear biometrics: A novel approach”, Journal of Intelligent systems, De_Gruiter publishing, advanced online publication, 2015.

[12] Egor P Popov, “Engineering mechanics of solids, Easter economy edition, 2nd edition 1998”.

[13] B Arbab Zavar, Mark S Nixon, “Model based analysis for ear biometric”, Computer Vision and Image understanding, vol 115, issue 4, P 487-502, 2011.

[14] Braja M. Das, Paul C. Hassler, “Statics and Mechanics of materials”, Prentice Hall, 1988.

[15] Anna Haugh, “Similarity measures for text Document clustering”, NZCSRSC 2008,April 2008,Christchurch,New Zealand.

[16] M. Steinbach, G. Karypis, and V. Kumar, “ A comparison of document clustering techniques” , In KDD Workshop on Text Mining, 2000.

[17] B. Larsen and C. Aone, “Fast and effective text mining using linear-time document clustering” , In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

[18] Thomas A Runkler, Data Analytics: Models and Algorithms for Intelligent data Analysis, Springer, 2012.

[19] Meghana Katri, Cosine similarity for the temporal dynamic web data, International Journal of Computer Science and Engineering Technology, Vol 3, No 8, 2012, pp315-318.

[20] Suphatkit Niwattanakul,Jatsada Singthongchai,Ekkaachai Naenudorn and Supachanun Wanapu , “using of Jaccard coefficient for keywords similarity”, Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong.

[21] Reza Bosagh Zadeh and Ashish Goel , “Dimension Independent Similarity Computation”, Journal of Machine Learning Research (2012) Submitted 6/12.

[22] GhanaShyam Nath Nayakam, “STUDY OF SIMILARITY COEFFICIENTS USING MAPREDUCE PROGRAMMING MODEL”, A Paper Submitted to the Graduate Faculty of the North Dakota State University of Agriculture and Applied Science

[23] Y. Zhao and G. Karypis, “Evaluation of hierarchical clustering algorithms for document datasets ”, In Proceedings of the International Conference on Information and Knowledge Management, 2002.