# A Comparative Study on Clustering Techniques for Urdu Ligatures in Nastaliq Font

Safia Shabbir[1], Nizwa Javed[2], Imran Siddiqi[1], Khurram Khurshid[2]
[1]Department of Computer Science,
Bahria University
Islamabad 44000, Pakistan
[2]Department of Electrical Engineering,
Institute of Space Technology,
Islamabad 44000, Pakistan
safiashabbir87@gmail.com, niz.jvd@gmail.com,
imran.siddiqi@gmail.com, khurram.khurshid@ist.edu.pk

***Abstract*—Clustering is a pivotal step in any Optical Character Recognition (OCR) or Word Spotting system. It serves as a base for the classification and indexing of different words or characters depending upon the level of segmentation. Various clustering methodologies have been applied by different researchers on Latin script based document images. However for Urdu language, which belongs to the family of Arabic and Persian, clustering based indexing systems have not been extensively researched. In this paper, we present a comprehensive study of various known clustering techniques applied on printed Urdu Document Images. The images are segmented into ligatures or partial words and then they are grouped together using different clustering methods. Performance of these methods is evaluated using Calinski-Harabasz, Davis-Bouldin and Dunn indexes.**

**Keywords: Document Image Analysis, Oriental script, Urdu Ligatures, Clustering, Word Spotting, OCR**

## I. Introduction

A lot of literature available in printed or handwritten format needs to be digitizes for future use. Where digitizing can be in the following forms: first is generating an editable document and the other is making an index file that help you find the scanned images of documents. Optical character recognition system caters the first form and the later is handled in a word spotting system. But in both these system we need to do clustering of characters or partial words for classification and making index file. Clustering is the process of grouping similar items in one cluster and dissimilar items in different clusters. Many OCR's and word spotting system have been proposed for like English, German, French, Arabic, Chinese etc. Only few researchers attempted to work on Cursive scripts like Urdu, Sindhi and Pashto which are spoken and used in South Asia specially in Pakistan.

A huge amount of literature is available for URDU in printed form which includes historic documents, poetry, novels, autobiographies etc. Urdu us a cursive script in which ligatures/ partial words are formed by joining different characters together. The composition of word and ligatures and its corresponding characters are shown in Figure 1.

The phenomenon of Grouping similar items on the basis of some features is termed as Clustering. Different algorithms like kmeans, Self Organizing map (SOM), hierarchical, sequential clustering have been used for different domains such as face recognition [1], script recognition and writer identification [2], word spotting [3], [4], information retrieval system [5], medical image segmentation [6]–[8], customer segmentation [9] etc. Clustering of Urdu ligatures/sub-words is not researched yet as separate problem. Although some work on Urdu OCR and word spotting exists which involves clustering as an intermediate step but those studies does not fully explain their clustering algorithms.

This particular study presents the comparative analysis of Kmeans, SOM and hierarchical clustering for Urdu ligatures. These clustering algorithms were evaluated on the basis of standard evaluation metrics like Calinski-Harabasz, Davies-Bouldin and Dunn index. The layout of the paper is as follows. The next section includes the review of work done on comparison of clustering techniques for word spotting and OCR and other related domains for Arabic, Farsi and other languages. Section III briefly explains the feature extraction and section IV present overview of clustering methodologies. Section V represents the mathematical ground for evaluation metrics for measuring clustering goodness. experiments and results are elaborated in section VI and section VII concludes the study.

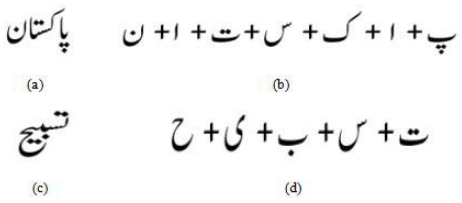


Fig. 1. Urdu Word composition a) Urdu Word Pakistan b) Characters of Pakistan c) Urdu word Tasbih d) Characters of Tasbih (Image Source:[10])

## II. Previous Work

Many researchers have worked on clustering for different domains. Few of them tried to improve existing algorithms, some of them gave comparison of clustering methods. Some of the studies are reviewed in this section.

Vuokko Vuori and Jorma Laaksonen in [11] presented a comparison on clustering algorithms for handwritten digits. Four hierarchical algorithms TreeCIusr, MinSwap, and two variations of C-means algorithm (named here CMeans 1 and CMeans 2) were used for the comparison. Clustering algorithms were evaluated on standard Davies-Bouldin and Calinski-Harabasz indexes. The results obtained with the clustering indexes were not promising at all rather a good set of digit prototype is found using combination of these clustering algorithms.
Ritu Yadav and Sunila Godara in [12] an improved hierarchical clustering algorithm using Euclidean distance formula. The proposed algorithm was experimented on 530 istances of A, B, C, D, E and F characters from Letter Image Recognition dataset. The algorithm was implemented and compared with normal Euclidean distance hierarchical clustering on Waikato Environment for Knowledge Analysis (Weka Tool). The accuracy of character recognition was improved by 1% using weighted Euclidean distance. Measuring accuracy for clustering goodness is not a good idea rather the study could have used standard indexes.

H. Arab Yarmohammadi *et al.* in [13] presented a comparison of K-means and hierarchical clustering for Farsi sub-words recognition. Zoning and Local Binary Patterns (LBP) were used as features. The cosine and correlation distances metrics were used for selecting first matched clusters. A dataset developed by Hoda System Company was used which includes six and half million words after removal of repetitive sub-words. After comparison both clustering algorithms, it was revealed that hierarchical clustering with correlation distance metric yield better results. But in hierarchical clustering the data is non-uniformly distributed. So the authors suggested to use k-Means rather than hierarchical clustering. Moreover its was also concluded that zoning features are more suitable rather than LBP's.

V. A. Pavlov and D. S. Shalymov in [14] presented a new distance measuering metric which is called feature relation graph (FRG) for clustering Arabic handwritten text . Arabic text lines written by different writers were clustered using FRG. Features are extracted using Gabor ans XGabor filters. After extraction of features maximum difference between those features is computed. After this FRG is calculated which is an unweighted directed graph that can vertices that denotes features. the proposed metric yielded 99% results with k-means clustering.

Akanksha Gaur and Sunita Yadav in [15] compared K-means and Support vector Machine for recognition of Hindi characters. Binarization and applied on input image and it is re-sized to 70x50. each image is divided in to seven partitions and k-means clustering using Euclidean distance is performed on this image. For classification SVM is used for 140 characters as training and rest 290 images are used as test data. SVM with linear kernel reported 95.86% accuracy.

## III. Data set and Features Extraction

Data set used for this particular study is gathered by scanning images of Urdu book 'Zawiya' written by 'Ashfaq Ahmed'. Initially 15 pages are used as input and preprocessing is performed on these images. A sample input image is shown in Figure 2. First of all text lines are extracted using methodology proposed in [16]. Few samples of extracted text lines are shown in Figure 3. There were total text lines and all of them were correctly segmented which results in 100% segmentation rate. For ligature segmentation connected component analysis is applied and 17376 ligatures were obtained as a result. In Urdu a there are two types of ligatures: Primary and secondary. Primary ligatures are those which forms the main body and secondary ligatures are diacritics associated with ligatures. In our work secondary ligatures are also considers as separate entity. Few examples of Primary and secondary ligatures are shown in Figure 4.

Following set of features are computed for each ligature: averages and standard deviations of horizontal projection,vertical projection, upper profile, lower profile. Two Discrete Cosine Transform (DCT) features, two Gabor filter features and three Discrete Wavelet transform (DWT) features. These features are explained in [17]. Each ligature is represented by a vector of length 15. Theses feature vectors ae used as an input to clustering algorithms.

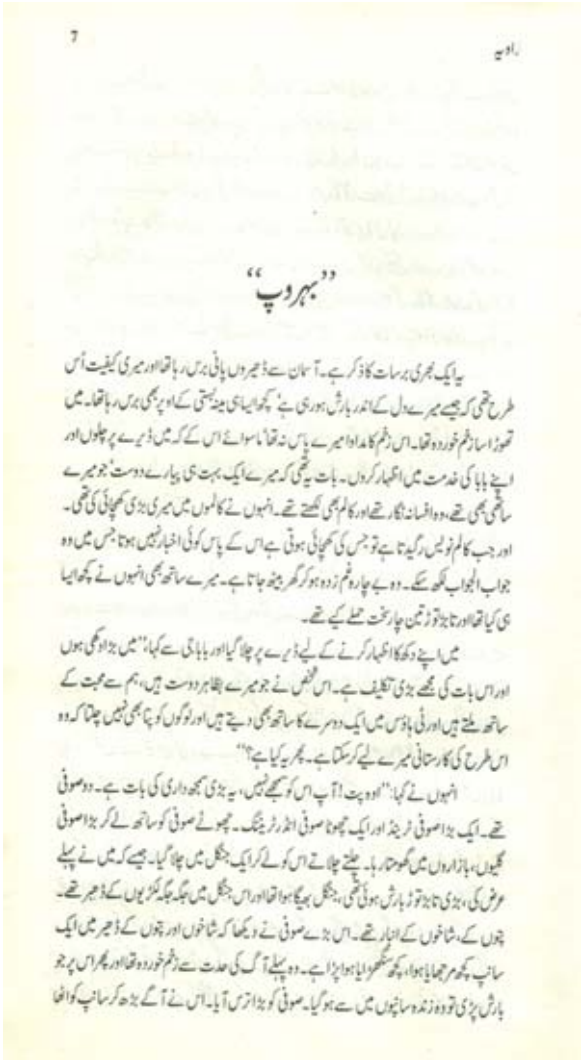

Fig. 2. Sample Data set image

Fig. 3. Samples of Extracted text lines

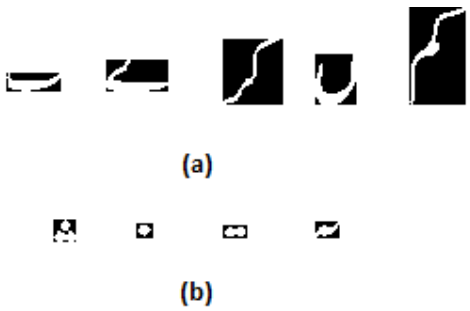


Fig. 4. (a) Primary ligatures (B) Secondary ligatures

## IV. Clustering Methods

Latest research on comparison among clustering techniques used traditional clustering algorithm in different domains. So in this work we have chosen conventional clustering methodologies i.e K-means , hierarchical and self organizing map clustering algorithms are employed for grouping Urdu ligatures. These three algorithms are briefly discussed in this section.

### A. K-Means clustering

K-means is a partition based clustering where each element is assigned to only one cluster. Each cluster is represented by its centroid and object is assigned to the cluster whose centroid has minimum distance from the object feature vector. In k-means clustering no of clusters i.e k must be given as an input to the algorithm.

### B. Hierarchical Clustering

There are two variations of hierarchical clustering: one is agglomerative and other is divisive. In agglomerative clustering each object or data point is considered as cluster and then those data point are merged to form clusters whereas in divisive clustering all points are considered to be part of one cluster and then they are separated into different clusters depending on their features. I this work we employed agglomerative clustering algorithm foe Urdu ligatures.

### C. Self Organizing Map (SOM) Clustering

Self Organizing Map is a type of artificial neural network which is trained using clustering to produce low dimensional representation of data. SOM use use competitive learning rather than error correction mechanism in contrast to other neural networks. It can have many dimensions like one dimension, two dimensions and three dimensions etc. It has many topologies formation such as: hexagonal, grid, random topology etc. In this particular study we have used one dimensional and hexagonal topology for clustering Urdu ligatures.

## V. Evaluation Metrics

There are several evaluation metrics that can be use for measring goodness of clustering agorithms such as: Silhouette index, Davies-Bouldin, Calinski-Harabasz, Dunn index, R-squared index, Hubert-Levin (C-index), Krzanowski-Lai index etc. Only three of them i.e Davies-Bouldin, Calinski-Harabasz and Dunn index are used in our study for comparison. Each lone of these are explained briefly in this section.

### A. Calinski-Harabasz Index

The Calinski-Harabasz criterion is also termed as variance ratio criterion (VRC). It is defined as:

$$VRC_k = \frac{SS_B}{SS_W} \times \frac{(N-k)}{(K-1)}$$

where $SS_B$ is the variance between the clusters, $SS_W$ is the variance within a cluster, k is the number of clusters, and N is the number of observations.

$SS_B$ is defined as:

$$SS_B = \sum_{i=1}^{k} n_i \left\| m_i - m \right\|^2$$

where k is the number of clusters, $m_i$ is the centroid of cluster i, m is the overall mean of the sample data, and $m_i - m$ Euclidean distance between the two vectors.

$SS_W$ is defined as:

$$SS_W = \sum_{i=1}^{k} \sum_{x \in c_i} \left\| x - m_i \right\|^2$$

where k is the number of clusters, x is a data point, $c_i$ is the ith cluster, $m_i$ is the centroid of cluster i.If the clusters are well separated then $SS_B$ is large and $SS_W$ is small. Larger value of Calinski-Harabasz index shows better clustering [18].

### B. Davies-Bouldin Index

The Davies-Bouldin criterion is based on a ratio of within-cluster and between-cluster distances. The mathematical explanation of Davies-Bouldin is as follows:

$$DB = \frac{1}{k} \sum_{i=1}^{k} max_{j \neq i} \{D_{i,j}\}$$

where $D_{i,j}$ is the within-to-between cluster distance ratio for the $i^{\text{th}}$ and $j^{\text{th}}$ clusters. Mathematically it can be expressed as:

$$D_{i,j} = \frac{\overline{d}_i + \overline{d}_j}{d_{i,j}}$$

$d_i$ is the average distance between each point in the $i^{\text{th}}$ cluster and the centroid of the $i^{\text{th}}$ cluster. $d_j$ is the average distance between each point in the $j^{\text{th}}$ cluster and the centroid of the $j^{\text{th}}$ cluster. $d_{i,j}$ is the Euclidean distance between the centroids of the $i^{\text{th}}$ and $j^{\text{th}}$ clusters. The optimal clustering solution has the smallest Davies-Bouldin index value [19].

### C. Dunn Index

Dunn index identifies well separated and dense clusters. It is the ratio between minimum inter-cluster distance and maximum intra-cluster distance which is mathematically expressed as:

$$D = \frac{min_{1 \leq i \leq j \leq n} d(i,j)}{max_{1 \leq k \leq n} d^{'}(k)}$$

The distance between cluster i and j is represented by $d(i,j)$ and $d^{'}(k)$ is the intra-cluster distance of cluster k. Since internal criterion seek clusters with high intra-cluster similarity and low inter-cluster similarity, algorithms that produce clusters with high Dunn index are more desirable [20].

## VI. Experiments and Results

Experiments were conducted using feature vector of all the ligatures obtained in ligature segmentation. K-means, Hierarchical and SOM clustering was applied for grouping ligatures into different numbers of clusters that is each clustering algorithm was used to partition the ligatures into 50, 60, 70, 80, 90 and 100 clusters.

K-means and hierarchical clustering and SOM algorithm requires number of clusters as an input argument. The data matrix is an input to K-means and Hierarchical clustering,in which each row represent one ligature. In contrast to k-means and hierarchical clustering,the data matrix required as an input for SOM represents one ligature in one column.

Execution time is an important aspect for clustering algorithms in order to assess whether they can group large data in less time or not. Table I shows the execution time comparison for k-means, hierarchical and SOM clustering algorithms in seconds. It is revealed that k-means is most efficient among other two algorithms.

The evaluation metrics explained in section V i.e Calinski- Harabasz (CH) , Davies-Bouldin (DB) and Dunn indexes are used for comparison purpose. The maximum values of CH and Dunn represents the optimal clustering solutions whereas in DB's minimum value represents optimal solution. The results of CH, DB and Dunn indexes for K-means clustering are presented in Table II. CH and DB gave optimal solution with 50 clusters whereas Dunn index gave optimal solution with 80 clusters. Optimal values are highlighted in the table.

Table III shows the results of CH, DB and Dunn obtained for hierarchical clustering. CH has maximum value at 50 no. of clusters, DB has minimum value for 50 no. of clusters and Dunn has maximum value for 50 no. of clusters as well. All three indexes gave optimal solution for hierarchical clustering with 50 clusters.

The results of CH, DB and Dunn indexes for SOM clustering are shown in Table IV. It was revealed that

TABLE I
COMPARISON OF EXECUTION TIME (MINUTES) FOR CLUSTERING ALGORITHMS

| Clusters/Algorithm | CK-Means | Hierarchical | SOM |
|---|---|---|---|
| **50** | 2.47 | 14.90 | 52.34 |
| **60** | 1.88 | 16.97 | 54.32 |
| **70** | 1.49 | 21.65 | 58.47 |
| **80** | 1.94 | 18.56 | 60 |
| **90** | 1.74 | 21.03 | 66.24 |
| **100** | 2.20 | 12.71 | 471.41 |

TABLE II
OVERVIEW OF THE RESULTS OBTAINED WITH INDEXES APPLIED TO THE K-MEANS ALGORITHM

| No. of clusters | CH | DB | Dunn |
|---|---|---|---|
| 50 | **13503** | **1.0841** | 0.0085 |
| 60 | 11407 | 1.1401 | 0.0015 |
| 70 | 11381 | 1.1694 | 0.00076 |
| 80 | 11007 | 1.1554 | **0.0126** |
| 90 | 10724 | 1.1834 | 0.0083 |
| 100 | 9920 | 1.2208 | 0.0082 |

TABLE III
OVERVIEW OF THE RESULTS OBTAINED WITH INDEXES APPLIED TO THE HIERARCHICAL CLUSTERING

| No. of clusters | CH | DB | Dunn |
|---|---|---|---|
| 50 | **741** | **0.7510** | **0.0766** |
| 60 | 616 | 0.8048 | 0.0742 |
| 70 | 528 | 0.8278 | 0.0711 |
| 80 | 463 | 0.8543 | 0.0724 |
| 90 | 416 | 0.8905 | 0.0747 |
| 100 | 417 | 0.8842 | 0.0738 |

TABLE IV
OVERVIEW OF THE RESULTS OBTAINED WITH INDEXES APPLIED TO THE SOM CLUSTERING

| No. of clusters | CH | DB | Dunn |
|---|---|---|---|
| 50 | **10343** | **1.1075** | 0.0020 |
| 60 | 7950 | 1.2209 | 0.0029 |
| 70 | 8238 | 1.1721 | 0.0025 |
| 80 | 8056 | 1.2015 | **0.0032** |
| 90 | 6730 | 1.2124 | 0.0013 |
| 100 | 7239 | 1.2413 | 0.0027 |

CH and DB has the optimal solution for 50 no. of clusters whereas Dunn index suggest that 80 no. of clusters is the optimal solution.

## VII. CONCLUSION

Clustering is an important step in any OCR or word spotting system. In this paper three clustering techniques K-means, hierarchical and SOM are compared. Results are compared on the basis of Calinski-Harabasz (CH) , Davies-Bouldin (DB) and Dunn indexes which are internal validity measures for clustering. The results shows that for all three clustering algorithms most of the indexes value gives optimal solution for 50 no. of clusters, So 50 clusters is the no. of groups in which 17376 clusters have been divided. On the basis of computation time K-means is found to be most efficient. In future this wok can be extended for clustering Urdu ligatures using state of the art deep learning concept. Current data set includes more occurrences of secondary ligatures as they are frequent in Urdu language, in future this work can be done for clustering only primary ligatures so that rare terms are also catered.

## ACKNOWLEDGEMENT

This research is funded by HEC Pakistan under the project 20-3853/R&D/HEC/14/126.

## REFERENCES

[1] J. Tao and Y. P. Tan, "Efficient clustering of face sequences with application to character-based movie browsing," in *2008 15th IEEE International Conference on Image Processing*, Oct 2008, pp. 1708–1711.

[2] S. Marinai, B. Miotti, and G. Soda, "Bag of characters and som clustering for script recognition and writer identification," in *2010 20th International Conference on Pattern Recognition*, Aug 2010, pp. 2182–2185.

[3] R. Hussain, H. A. Khan, I. Siddiqi, K. Khurshid, and A. Masood, "Keyword based information retrieval system for urdu document images." in *SITIS*. IEEE Computer Society, 2015, pp. 27–33.

[4] A. Abidi, I. Siddiqi, and K. Khurshid, "Towards searchable digital urdu libraries - a word spotting based retrieval approach." in *ICDAR*. IEEE Computer Society, 2011, pp. 1344–1348.

[5] M. R. Hussain, A. Masood, H. Khan, K. Khurshid, and I. Siddiqi, "Language independent keyword based information retrieval system of handwritten documents using svm classifier and converting words into shapes," *Pakistan Journal of Engineering and Applied Sciences*, vol. 0, no. 0, 2016.

[6] A. Martnez-Us, F. Pla, and P. Garca-Sevilla, *Unsupervised Image Segmentation Using a Hierarchical Clustering Selection Process*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 799–807.

[7] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image segmentation using k -means clustering algorithm and subtractive clustering algorithm," *Procedia Computer Science*, vol. 54, pp. 764 – 771, 2015.

[8] K. Shrivastava, N. Gupta, and N. Sharma, "Medical image segmentation using modified k means clustering," *International Journal of Computer Applications*, vol. 103, no. 16, pp. 12–16, 2014.

[9] W. Gao, G. Qian, and H. Xu, "Som clustering analysis for telecommunication customer segmentation," in *2009 International Conference on Management and Service Science*, Sept 2009, pp. 1–4.

[10] S. Naz, K. Hayat, M. I. Razzak, M. W. Anwar, S. A. Madani, and S. U. Khan, "The optical character recognition of urdu-like cursive scripts," *Pattern Recognition*, vol. 47, no. 3, pp. 1229–1248, 2013.

[11] V. Vuori and J. Laaksonen, "A comparison of techniques for automatic clustering of handwritten characters," in *16th International Conference on Pattern Recognition, ICPR 2002, Quebec, Canada, August 11-15, 2002.*, 2002, pp. 168–171.

[12] R. Yadav and S. Godara, "An improved hierarchical clustering technique for character recognition," *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, no. 7, pp. 1229–1248, September 2012.

[13] H. A. Yarmohammadi, A. A. Fard, and H. Khosravi, "Clustering low quality farsi sub-words for word recognition," in *2014 Iranian Conference on Intelligent Systems (ICIS)*, Feb 2014, pp. 1–5.

[14] V. A. Pavlov and D. S. Shalymov, "Arabic handwritten texts clusterization based on feature relation graph (frg)," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Aug 2015, pp. 941–945.
[15] A. Gaur and S. Yadav, "Handwritten hindi character recognition using k-means clustering and svm," in *2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services*, Jan 2015, pp. 65–70.
[16] I. S. Israr Ud Din, Zumra Malik and S. Khalid, "Line and ligature segmentation in printed urdu document images," *Journal of Applied Environmental and Biological Sciences*, vol. 6, no. 3, pp. 114–120, March 2016.
[17] I. S. Maria Siddiqui and K. Khurshid, "Feature extraction for cursive language document images using discrete cosine transform, discrete wavelet transform and gabor filter," in *1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, April 2017.
[18] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-Simulation and Computation*, vol. 3, no. 1, pp. 1–27, 1974.
[19] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 224–227, 1979.
[20] J. C. Dunn, "Well separated clusters and optimal fuzzy-partitions," *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974.