International Journal of Computer Applications Technology and Research
Volume 6–Issue 5, 213-223, 2017, ISSN:-2319–8656

Optimal Clustering Technique for Handwritten Nandinagari Character Recognition

Prathima Guruprasad
Research Scholar, UOM,
Dept. of CSE, NMIT,
Gollahalli, Yelahanka,
Bangalore, India

Prof. Dr. Jharna Majumdar
Sc. G DRDO (Retd.), Dean,
R&D, Prof. and Head, Dept. of CSE and Center for
Robotics Research, NMIT, Bangalore, India

Abstract: In this paper, an optimal clustering technique for handwritten Nandinagari character recognition is proposed. We compare two different corner detector mechanisms and compare and contrast various clustering approaches for handwritten Nandinagari characters. In this model, the key interest points on the images which are invariant to Scale, rotation, translation, illumination and occlusion are identified by choosing robust Scale Invariant Feature Transform method(SIFT) and Speeded Up Robust Feature (SURF) transform techniques. We then generate a dissimilarity matrix, which is in turn fed as an input for a set of clustering techniques like K Means, PAM (Partition Around Medoids) and Hierarchical Agglomerative clustering. Various cluster validity measures are used to assess the quality of clustering techniques with an intent to find a technique suitable for these rare characters. On a varied data set of over 1040 Handwritten Nandinagari characters, a careful analysis indicate this combinatorial approach used in a collaborative manner will aid in achieving good recognition accuracy. We find that Hierarchical clustering technique is most suitable for SIFT and SURF features as compared to K Means and PAM techniques.

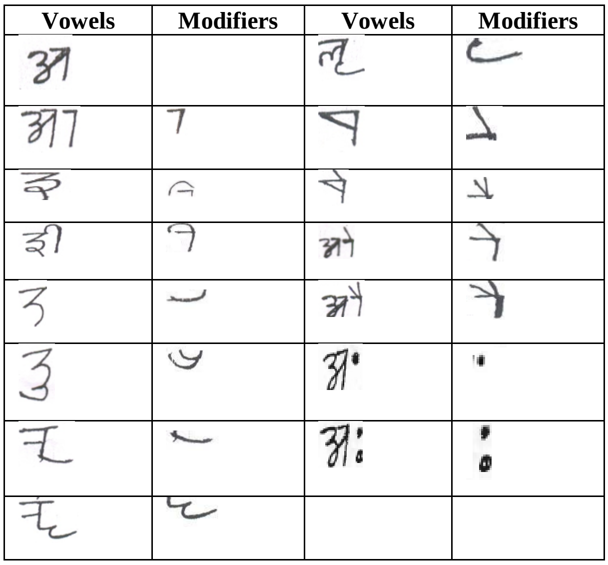
Keywords: Invariant Features, Scale Invariant Feature Transform, Speeded Up Robust Feature technique, Nandinagari Handwritten Character Recognition, Dissimilarity Matrix, Cluster measures, K Means, PAM, Hierarchical Agglomerative Clustering

1. INTRODUCTION

The awareness of very old scripts is valuable to historians, archaeologists and researchers of almost all branches of knowledge for enabling them to understand the treasure contained in ancient inscriptions and manuscripts [1]. Nandinagari is a Brahmi-based script that was existing in India between the 8th and 19th centuries. This is used as writing style in Sanskrit especially in southern part of India. Nandinagari script is older version of present day Devanagari script. But there are some similarities between Nandinagari and Devanagari in terms of their character set, glyphic representation and structure. However, Nandinagari differs from Devanagari in the shapes of character glyphs, absence of headline. There are several styles of Nandinagari, which are to be treated as variant forms of the script. Sri Acharya Madhwa of the 13th century, a spiritual Leader who founded the Dvaita school of Vedanta has hundreds of manuscripts written in Nandinagari on the Palm leaves.

Nandinagari script is available only in manuscript form hence it lacks the necessary sophistication and consistency. There are innumerable manuscripts covering vast areas of knowledge, such as Vedas, philosophy, religion, science and arts preserved in the manuscript libraries in digital form. Today though Nandinagari script is no longer in trend, the scholars of Sanskrit literature cannot be ignorant of this script. Nandinagari character set has 15 vowels and 37 consonants, 52 characters as shown in Table 1 and Table 2. We face many challenges to interpret handwritten Nandinagari characters such as handwriting variations by same or different people with wide variability of writing styles. Further, these documents are not available in Printed Format and only handwritten scripts are available. Absence of any other published research methods using these rare characters makes if more challenging. Nandinagari Optical Character Recognition (OCR) is not available to date. Therefore, we need to extract invariant features of these handwritten characters to get good recognition accuracy.

Table 1. Nandinagari Vowels and Modifiers



In this paper we extract features using Scale Invariant Feature Transform (SIFT) [2] and Speeded Up Robust Feature (SURF) transform techniques [7]. The SIFT and SURF features are local and based on the appearance of the object and are invariant to different sizes and orientations. They are also robust to changes in illumination, noise and highly distinctive with low probability of mismatch. From these features, a dissimilarity matrix is computed. Then this is given as an input to different clustering techniques to group similar characters. The set of clustering mechanisms identified for these characters are K Means, PAM and Hierarchical agglomerative clustering

technique. The performance of these techniques are compared and best method for SIFT and SURF features is identified.

Table 2. Nandinagari Consonants

2. RELATED WORK

The scale and variety of applications using SIFT [3][4][5][6] and SURF[9] is discussed in many papers on pattern recognition. The robustness of SIFT and its comparison with SURF algorithm is also discussed in some papers [8][9]. The recognition of multiple type of words including Devanagari using Visual bag of words is discussed using SIFT algorithm [11]. An attempt to classify human faces using SIFT and hierarchical clustering approach is also introduced [12].Clustering algorithms like K Means and K Medoids and their performance is also discussed [14].Different cluster measures to evaluate the performance of cluster methods are also discussed [15].

3. METHODOLOGY

Handwritten Nandinagari character database is created manually as standard dataset is not available. For a set of 52 vowels and consonants in Nandinagari, with an average of 5 different variations over the format of representation(jpg or png), size(256X256, 384X384, 512X512, 640X640), degree of rotation(0, 45, 90, 135 and 180 degree) and translation(positive or negative offset of 15 pixels), a database of 1040 characters is prepared. The proposed architecture shown in Fig. 1 consists of following steps:

1. In the first step, all the characters in the database are scanned.
2. In the pre-processing step, we convert these images into their grayscale form.
3. Interest points from the input image are extracted using the Scale invariant feature transform (SIFT) technique. From each point, 128 feature descriptors are extracted which are invariant to scale, rotation and illumination. Similarly, features are also extracted using Speeded Up Robust Feature (SURF) transform technique. 64 feature descriptors are generated from each candidate points.
4. For each image in the database, the number of match points are found with every other image and vice versa.
5. The maximum number of match points is computed by considering number of match points and N X N match matrix is generated.
6. The dissimilarity ratio is now computed using the following formula

Eij = Eji = {100 * (1 - nMax / nMin) }, where nMax = maximum number of match points in either directions and nMin = minimum number of key points in either direction

7. The SIFT and SURF features dissimilarity matrix is fed as input for different clustering techniques to group similar handwritten Nandinagari characters together.
8. The best-suited clustering technique for SIFT and SURF features is identified by analysing the performance using cluster measures.

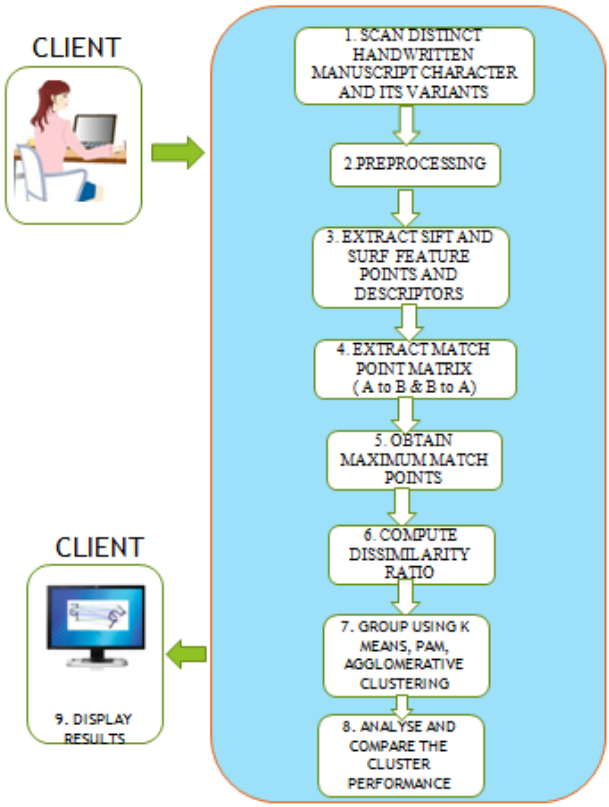


Figure 1. Proposed Model Architecture

3.1 Clustering

Clustering is the process of grouping a set of similar objects into same clusters and dissimilar objects in other clusters. Three prominent approaches are taken for analysis and comparison here. They are K Means, PAM and Agglomerative Hierarchical Clustering.

3.1.1 K Means Clustering

K-means clustering algorithm uses an iterative refinement approach. Here we partition of the characters into k clusters, such that the characters in a cluster are more similar to each other than to characters in different clusters [14].This is based on the Euclidean distance measure, we calculate the new mean

which is the centroid of the clusters and assign nearest points and this process is continued until the cluster centres remains unchanged.

3.1.2 PAM Clustering

This method chooses a character from all characters in the dataset as medoids of a cluster i.e., a cluster centre, in contrast to the K-Means method, which selects a random value as the centre of the cluster. The objective is to minimize the average dissimilarity of characters to their closest selected character. This method starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering [15]. The PAM method is more robust to noise and outliers, compared to the K-means method.

3.1.3 Agglomerative Clustering

Agglomerative Hierarchical Clustering is a bottom up approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up in the hierarchy [13].The result of the hierarchical methods is a dendrogram, representing the nested grouping of objects. There are different methods for agglomeration such as single, complete, average methods. In this paper, we have used the average linkage method as an algorithm for this approach. This is better than the K Means and PAM approaches since it automatically detects the number of clusters.

3.2 Cluster Validation Measures

Choosing appropriate clustering method for a given dataset is a very challenging task. So different clustering measures are considered to validate the clustering results. It is helpful to choose best clustering technique for a specific application. Here we validate the results using two categories of measures such as internal and stability validation measures. Internal measure use the fundamental information in the data to evaluate the quality of the clustering. Stability measure evaluate the consistency of the clustering results [16].

3.2.1 Internal Measures

For internal measures, three measures are considered such as Connectivity, Silhouette width and Dunn Index. Connectivity is used to measure the connected component, which relates to what extent items are placed in the same cluster as their nearest neighbours in the data space. In the second measurement approach, the silhouette value measures the degree of confidence in the clustering assignment of a particular item. This value ranging between -1 to 1 need to be maximized. However, in the third measurement approach, the Dunn index indicates the ratio of the smallest distance between items not in the same cluster to the largest intra-cluster distance. The Dunn index has a value between zero and one, and need to be maximized.

3.2.2 Stability Measures

The stability measure compare the results from clustering based on the original data set to clustering based on deleting one column at a time. These measures work well if the data are highly correlated. The stability measures considered here are the average proportion of non-overlap (APN), the average distance (AD), the average distance between means (ADM), and the figure of merit (FOM). The APN measures the average proportion of observations not placed in the same cluster by clustering based on the original data and clustering based on the data with a single column removed. The AD measure computes the average distance between observations placed in the same cluster by clustering based on the original data and clustering based on the data with a single column removed. The ADM measure computes the average distance between cluster centres for observations placed in the same cluster by clustering based on the original data and clustering based on the data with a single column removed. The FOM measures the average intra-cluster variance of the observations in the removed column, where the clustering is based on the remaining samples. This estimates the mean error using predictions based on the cluster averages. The APN has the value between 0 and 1 and with values close to zero corresponds to highly consistent clustering results. Remaining measures have values between zero and ∞and smaller values are favoured for better clustering performance.

4. EXPERIMENTAL RESULTS

The results are obtained for various stages of character recognition. The samples of images of different size 256 X 256, 384 X 384, 512 X 512,640 X 640, different orientation angles 0o, 45o, 90o, 135o, 180o are taken. This forms a 1040 character in the database. All the 1040 characters are considered for computation and for the sake of depicting the results of cluster formation, we take a subset of 16 distinct characters from this set.

4.1 Cluster using SIFT Features

For K-means clustering approach, the parameter to be set prior to clustering is the number of clusters. The optimal number of clusters i.e. 14 for SIFT features is derived using Elbow method is as shown in Fig 2. The clusters obtained using this technique is indicated in Appendix1. As seen in this figure, the instances are misclassified and hence would yield a low accuracy rate. For PAM the optimal number of clusters need to be mentioned but the partition done around the medoids and this is better compared to K Means approach. The cluster results using SIFT features are as shown in Appendix 2.

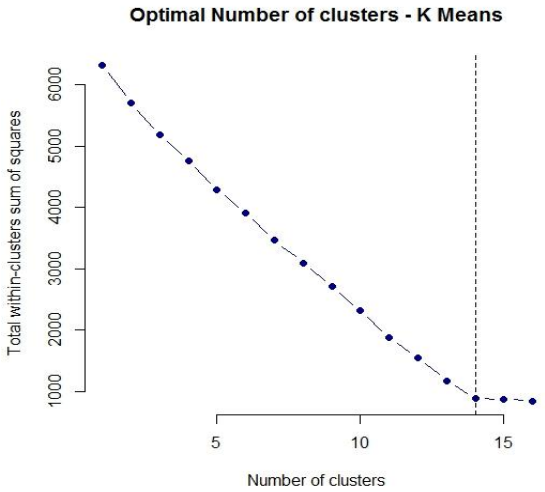


Figure 1. Optimal Number of Clusters (14) using K-means for SIFT features

The optimal number of clusters for PAM is same as K Means method as shown in Fig 3.The dendogram after hierarchical clustering using SIFT features for sample characters partitioned automatically into 16 clusters is as shown in Appendix 3.

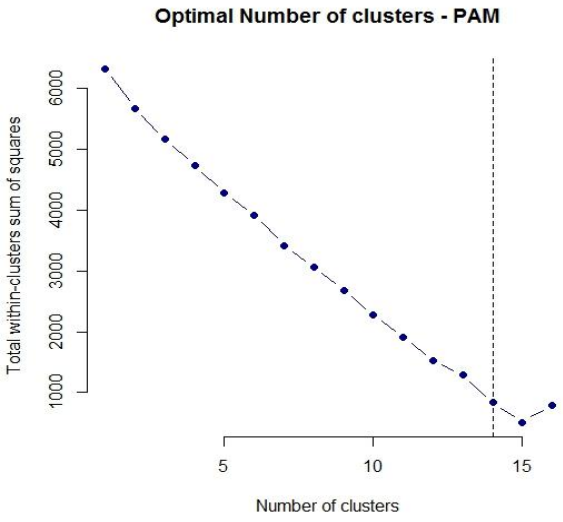


Figure 3. Optimal Number of Clusters (14) using PAM for SIFT features

4.1.1 Cluster Validation Measures for SIFT Features

The internal validation measures are the Connectivity, Silhouette width, Dunn Index derived for three different clustering techniques, K Means, PAM, and agglomerative hierarchical clustering techniques using SIFT features. The clustering validation results are analysed and the optimal score for these three measures as shown in Table 3. For internal measures using SIFT features, hierarchical clustering with two clusters performs better for connectivity measures. For Dunn Index and Silhouette Width, hierarchical clustering with fourteen clusters performs better. For good clustering, the connectivity is minimized, while both the Dunn index and the silhouette width is maximized. So from table 3 it appears that hierarchical clustering performs better compared to the other clustering techniques for each internal validation measure.

Table 3. Internal and Stability cluster validation measures for SIFT Features

Internal Measures			
Measures	**Value**	**Cluster Method**	**No. of Clusters**
Connectivity	8.5079	Hierarchical	2
Dunn Index	0.9191	Hierarchical	14
Silhouette Width	0.7140	Hierarchical	14
Stability Measures			
Measures	**Value**	**Cluster Method**	**No. of Clusters**
APN	0.0195	Hierarchical	14
AD	61.1687	Hierarchical	14
ADM	7.6345	Hierarchical	14
FOM	8.8808	Hierarchical	14

The graphical representation of the connectivity, Dunn index, and Silhouette Width measures are as shown in Fig.4 to 6.

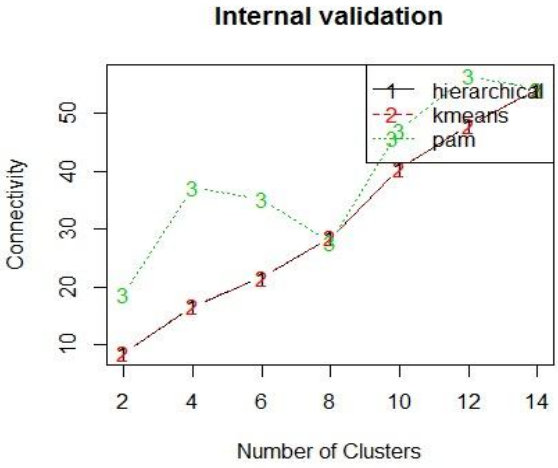


Figure 4. Graphical representation of the connectivity internal measure using SIFT Features

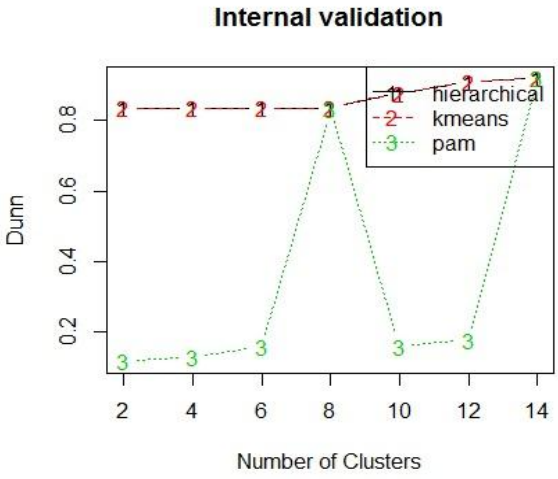


Figure 5. Graphical representation of the Dunn index internal measure using SIFT Features

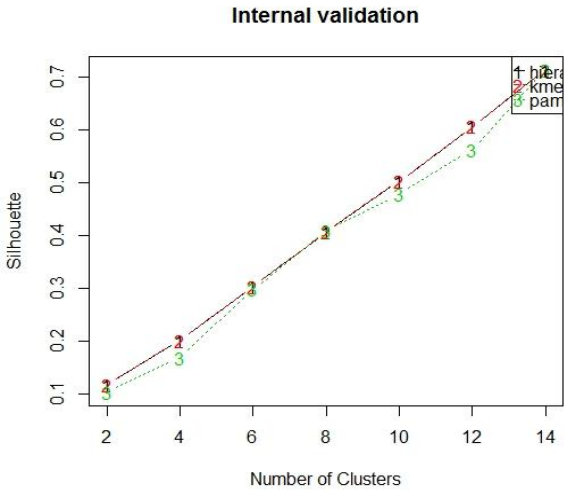


Figure 6. Graphical representation of the Silhouette Width internal measure using SIFT Features

The stability measures for K Means, PAM and agglomerative hierarchical clustering techniques using SIFT features are computed. The optimal scores of the measures such as APN, AD, ADM, and FOM are as shown in Table 3. For better clustering results the measures are minimized. From the table 3, for these measures, hierarchical clustering with fourteen clusters gives the best score. The graphical representation of the stability measures for SIFT features such as APN, AD, ADM and FOM as shown in Fig.7 to Fig. 10.

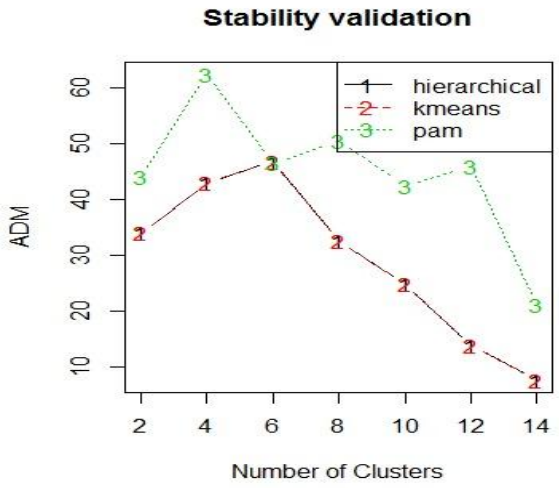


Figure 7. Graphical representation of the ADM stability measure using SIFT Features

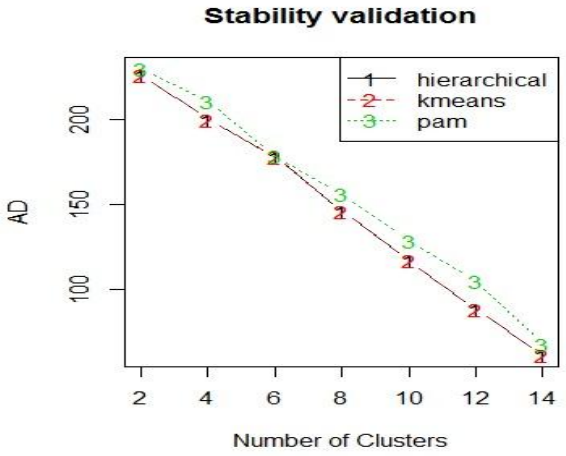


Figure 8. Graphical representation of the AD stability measure using SIFT Features

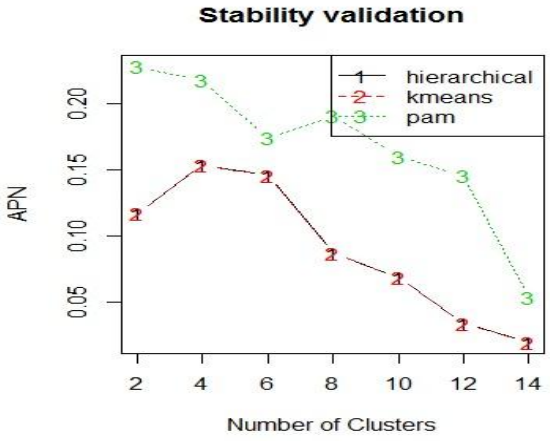


Figure 9. Graphical representation of the APN stability measure using SIFT Features

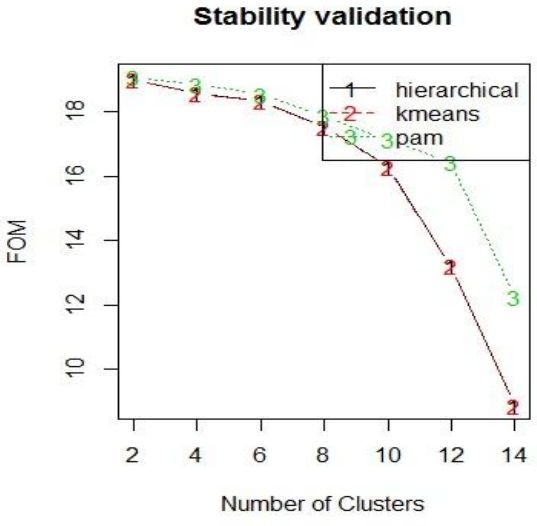


Figure 10. Graphical representation of the FOM stability measure using SIFT Features

4.2 Clustering using SURF Features

For K-means clustering technique, the number of clusters are decided by Elbow method for SURF features. The k value is found as 16 as shown in Fig 11. The SURF features grouped together using this approach is as shown in Appendix 4. The misclassification rate is more in K Means method. For PAM the optimal number of clusters is generated using Elbow method and as shown in Fig 12.

PAM is better compared to K Means approach because partition is done around the medoids which leads to low error rate. The cluster results using SURF features is as shown in Appendix 5.

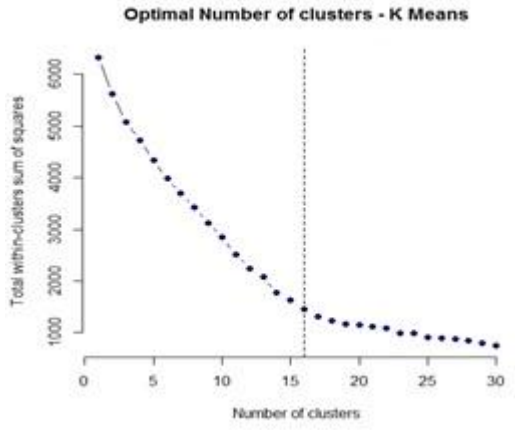


Figure 11. Optimal Number of Clusters (16) using K-Means for SURF features

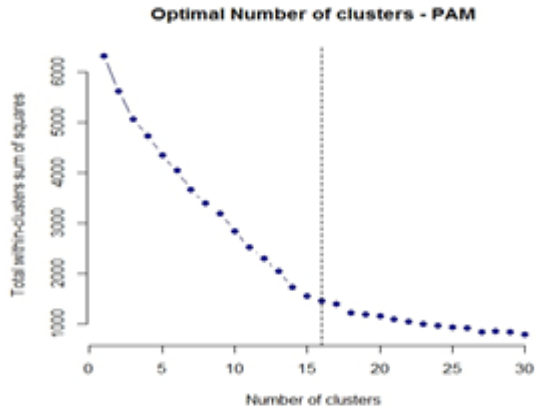


Fig. 12. Optimal Number of Clusters (16) using PAM for SURF features

The dendogram of hierarchical clustering using SURF features for sample characters partitioned automatically into 16 clusters is as shown in Appendix 6.

4.2.1 Cluster Validation Measures for SURF Features

The internal and stability cluster validation measures for SURF features is used to evaluate the results of K Means, PAM and agglomerative Hierarchical clustering methods.

The analysis is as shown in table 4 for different cluster measures. Internal clustering validation, which use the internal information of the clustering process to evaluate the efficiency of a clustering method. It can be seen that for SURF features among three clustering methods, hierarchical clustering with 2 clusters performs better for Connectivity and with 16 clusters for Dunn Index and Silhouette Width.

Clustering stability validation evaluates the consistency of a clustering result by comparing it with the clusters obtained after each column is removed, one at a time. It is analysed that for SURF features, Hierarchical clustering with 16 clusters proved to be better for APN, AD, ADM, FOM stability measures.

Table 4. Internal and Stability cluster validation measures for SIFT Features

Internal Measures			
Measures	**Value**	**Cluster Method**	**No. of Clusters**
Connectivity	8.5079	Hierarchical	2
Dunn Index	3.0797	Hierarchical	16
Silhouette Width	0.8153	Hierarchical	16
Stability Measures			
Measures	**Value**	**Cluster Method**	**No. of Clusters**
APN	0.0000	Hierarchical	16
AD	34.2118	Hierarchical	16
ADM	0.0000	Hierarchical	16
FOM	3.4201	Hierarchical	16

The corresponding graphical representation of these measures as shown in Fig.13 to 19.

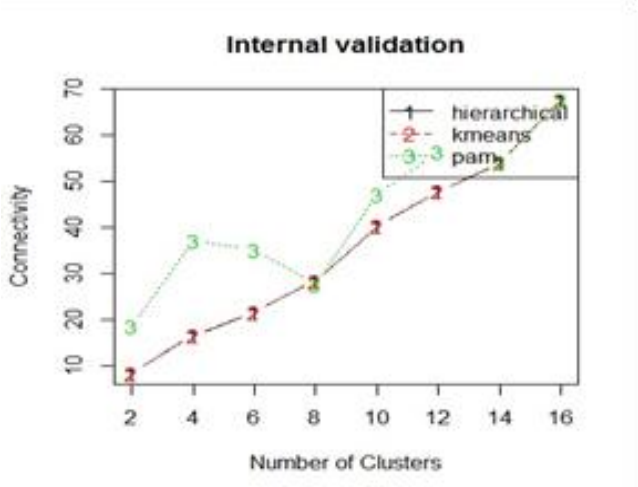


Figure 13. Connectivity internal measure for SURF Features

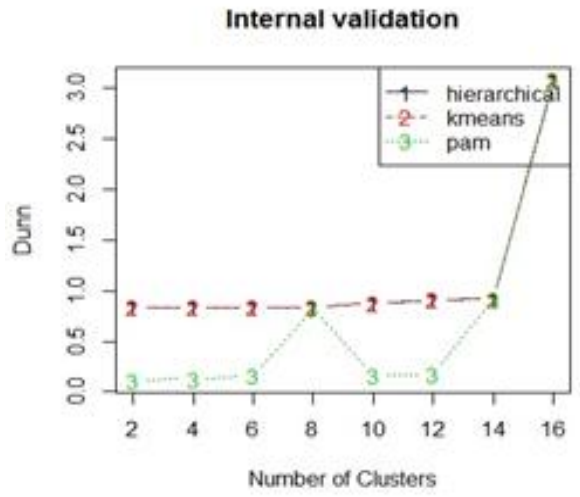


Figure 14. Dunn Index internal measure for SURF Features

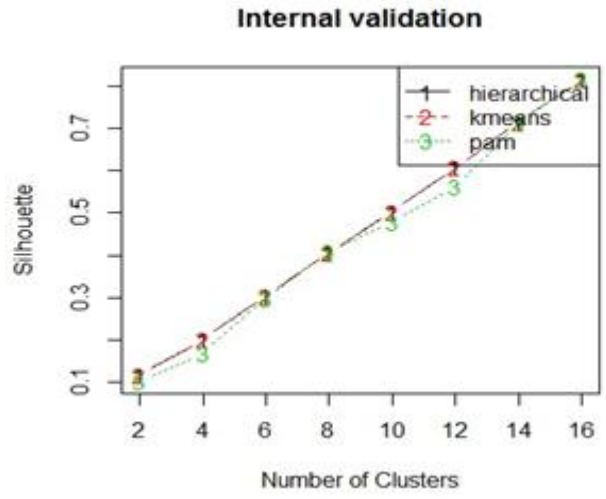


Figure 15. Silhouette Width internal measure for SURF Features

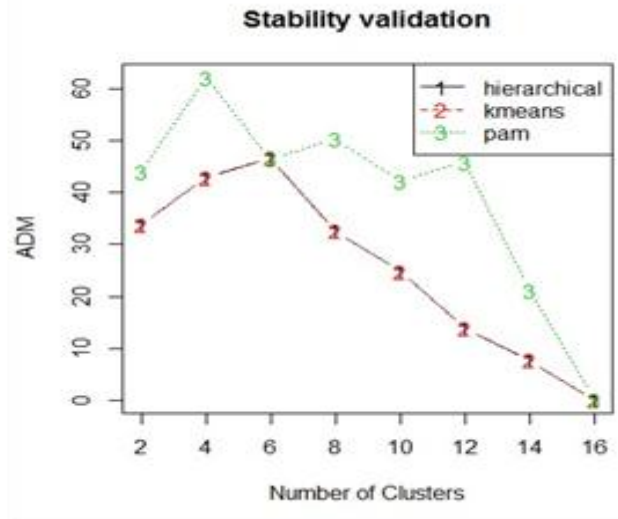


Fig. 16: ADM stability measure for SURF Features

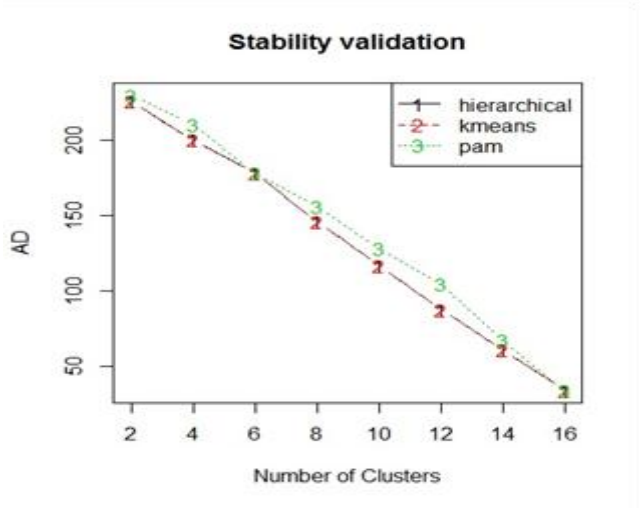


Fig. 17: AD stability measure for SURF Features

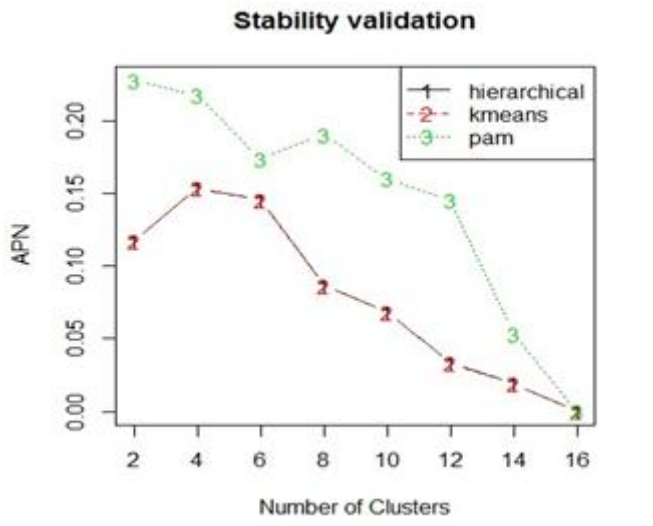


Fig. 18: APN stability measure for SURF Features

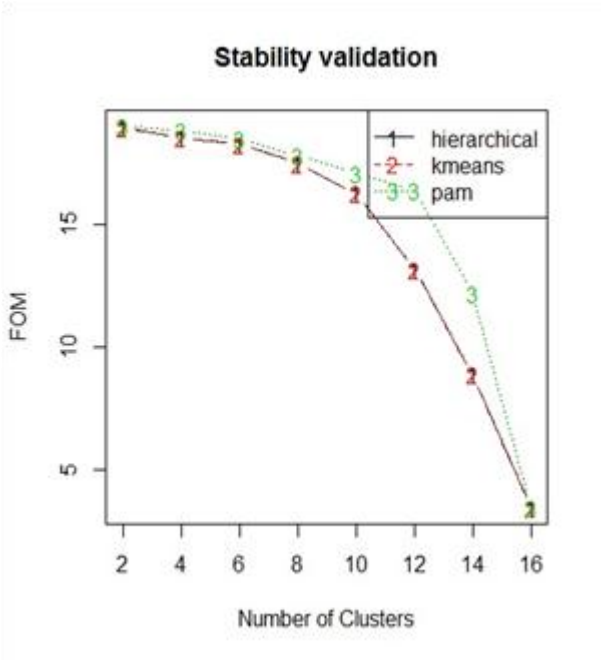


Fig. 19: FOM stability measure for SURF Features

5. CONCLUSION

The proposed Nandinagari character retrieval system based on data visualization method and is highly scalable. The SIFT and SURF methods detect the interest points and derives feature descriptors. This approach requires no or minimal pre-processing of images and still can identify images in varying states of occlusion. Our main aim is to provide efficient and robust descriptors which are then used to compute dissimilarity matrix. SIFT descriptors are more robust compared to SURF descriptors. But computation time for SURF is less compared to SIFT method. Then dissimilarity matrix of these descriptors are subjected to different clustering approaches to group similar handwritten Nandinagari characters together. Prerequisite for K-Means and PAM is to specify the number of clusters. Performance of PAM is better compared to K Means. Agglomerative clustering method is more suitable for both SIFT and SURF descriptors. Further we can explore the performance of these descriptors using wide variety of clustering techniques.

6. REFERENCES

[1] P. Visalakshi, "Nandinagari Script", DLA Publication, First Edition, 2003.

[2] D. G. Lowe, “Distinctive image features from scale-invariant key points”, IJCV, vol. 60, no. 2, pp. 91–110, 2004.

[3] Mortensen, E.N., Deng, H., Shapiro, L., "A SIFT descriptor with global context", In Computer Vision and Pattern Recognition (CVPR 2005), 20-25,IEEE, Vol. 1, 184-190, June 2005..

[4] Ives Rey-Otero, Jean-Michel Morel and Mauricio Delbarcio, “An analysis of Scale-space sampling in SIFT”, ICIP, 2014.

[5] Yishu Shi, Feng Xu, Feng-Xiang Ge Yishu Shi, Feng Xu, Feng-Xiang Ge “SIFT-type Descriptors for Sparse-Representation Based Classification”, 10th International Conference on Natural Computation IEEE 2014.

[6] D. G. Lowe, “Object recognition from local scale-invariant features,” ICCV, 1999.

[7] Baofeng Zhang, Yingkui Jiao Zhijun Ma, Yongchen Li, Junchao Zhu,"An Efficient Image Matching Method Using Speed Up Robust Features", International Conference on Mechatronics and Automation IEEE 2014.

[8] Dong Hui, Han Dian Yuan, "Research of Image Matching Algorithm Based on SURF Features", International Conference on Computer Science and Information Processing (CSIP) , IEEE 2012.

[9] Goh K M, Abu-Bakar S A R, Mokji M M, et al., " Implementation and Application of Orientation Correction on SIFT and SURF Matching",IJASCA, vol.5, no.3, 2013.

[10] Martin A. Fischler and Robert C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” in Journal on Communications of the ACM, ACM Press, NY Vol 24, No. 6, pages 381 to 395, 1981.

[11] Ravi Shekhar, C.V. Jawahar, “Word Image Retrieval using Bag of Visual Words” IEEE, DOI 10.1109/DAS.2012.96, 2012.

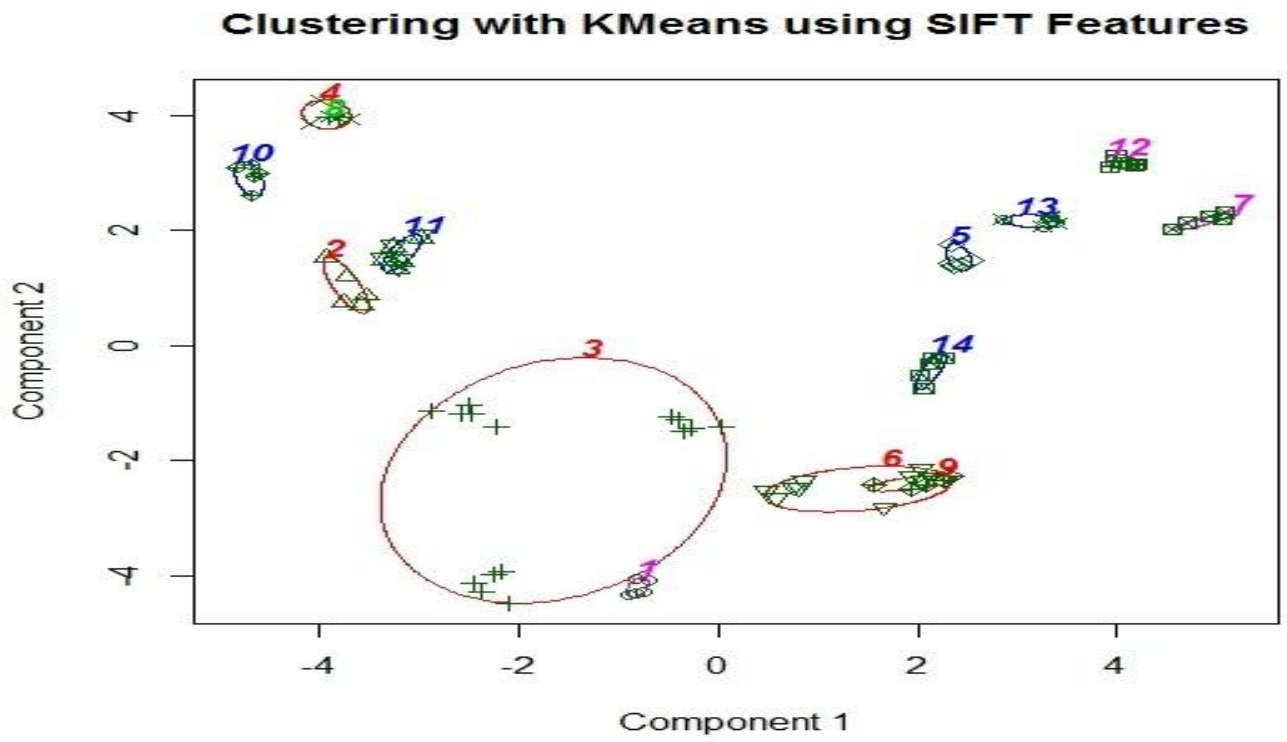
[12] Panagiotis Antonopoulos, Nikos Nikolaidis and Ioannis Pitas, “Hierarchical Face Clustering Using Sift Image Features”, Computational Intelligence in Image and Signal Processing, IEEE Symposium on April 2007.

[13] Akanksha Gaur, Sunita Yadav, "Handwritten Hindi Character Recognition using KMeans Clustering and SVM", 2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services, IEEE 2015.

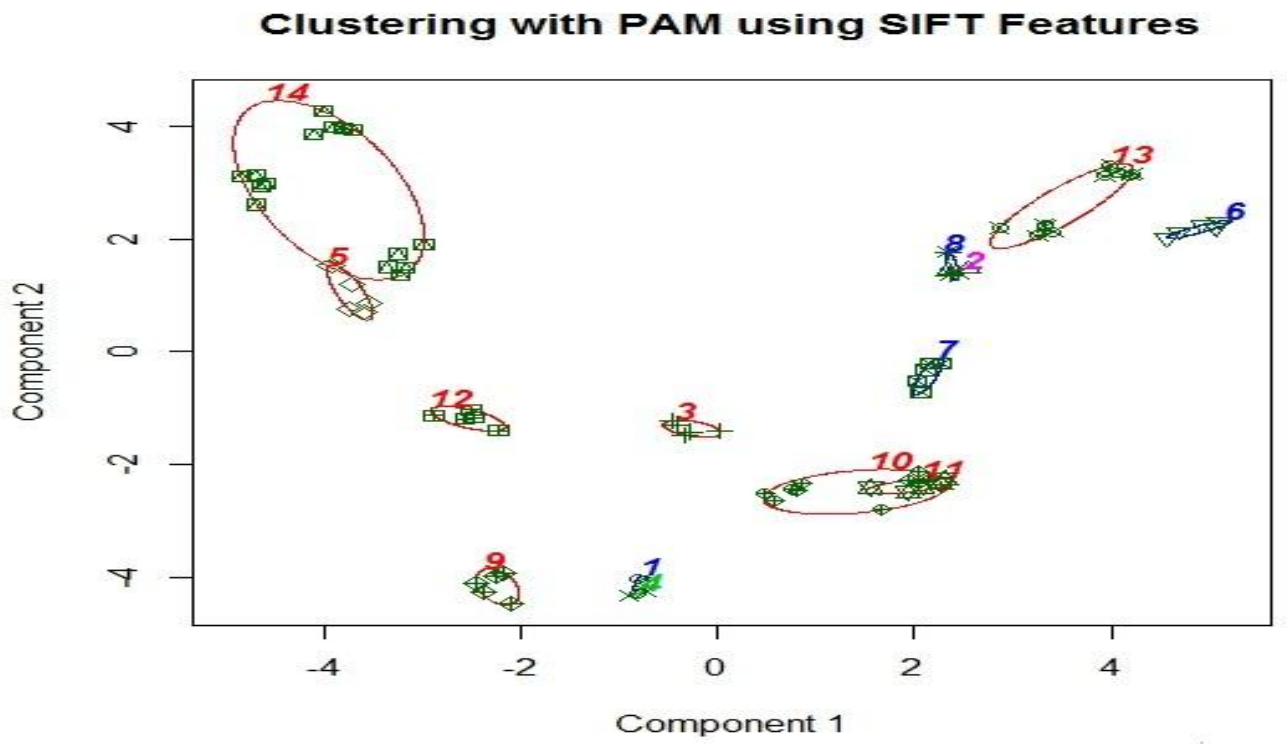
[14] Hae-Sang Park, Jong-Seok Lee, Chi-Hyuck Jun , " A K-means-like Algorithm for K-medoids Clustering and Its Performance ",

[15] Oliver Kirkland, Beatriz De La Iglesia, " Experimental Evaluation of Cluster Quality Measures", IEEE, 2013

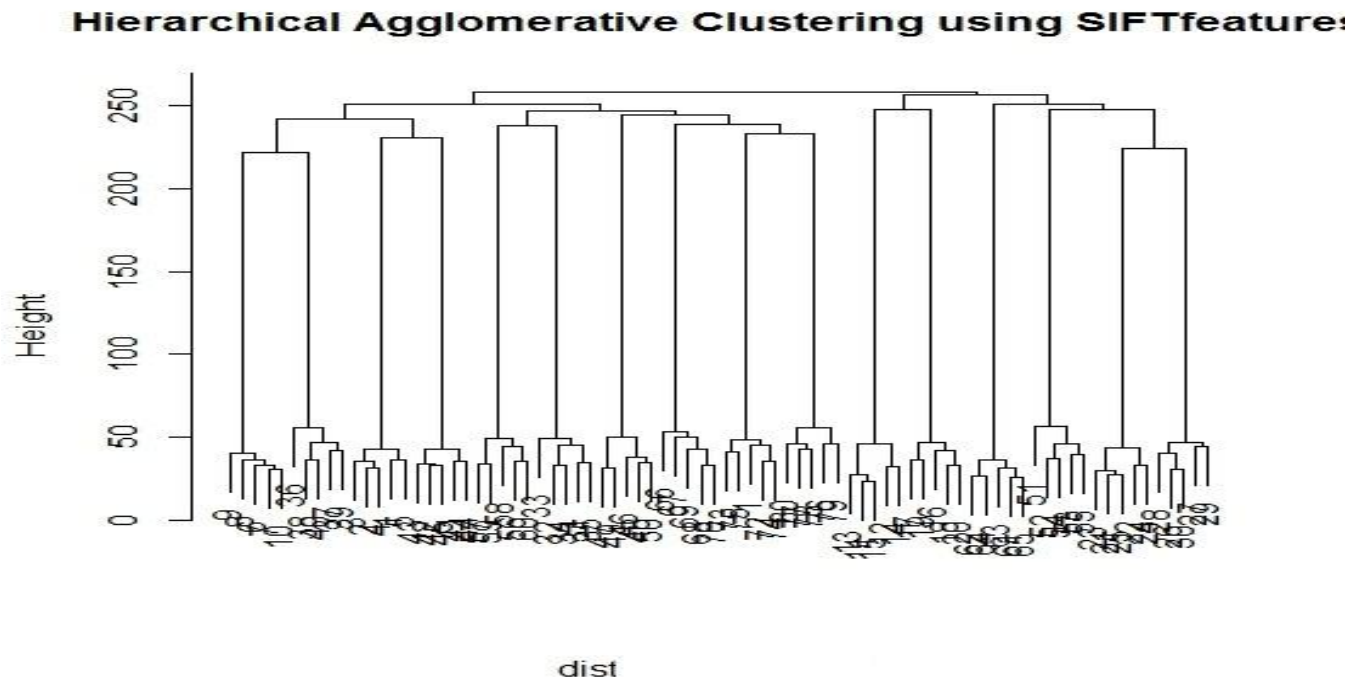
[16] Guy Brock, Vasyl Pihur, Susmita Datta, Somnath Datta," clValid: An R Package for Cluster Validation " Journal of Statistical Software, Vol 25, Issue 4, March 2008



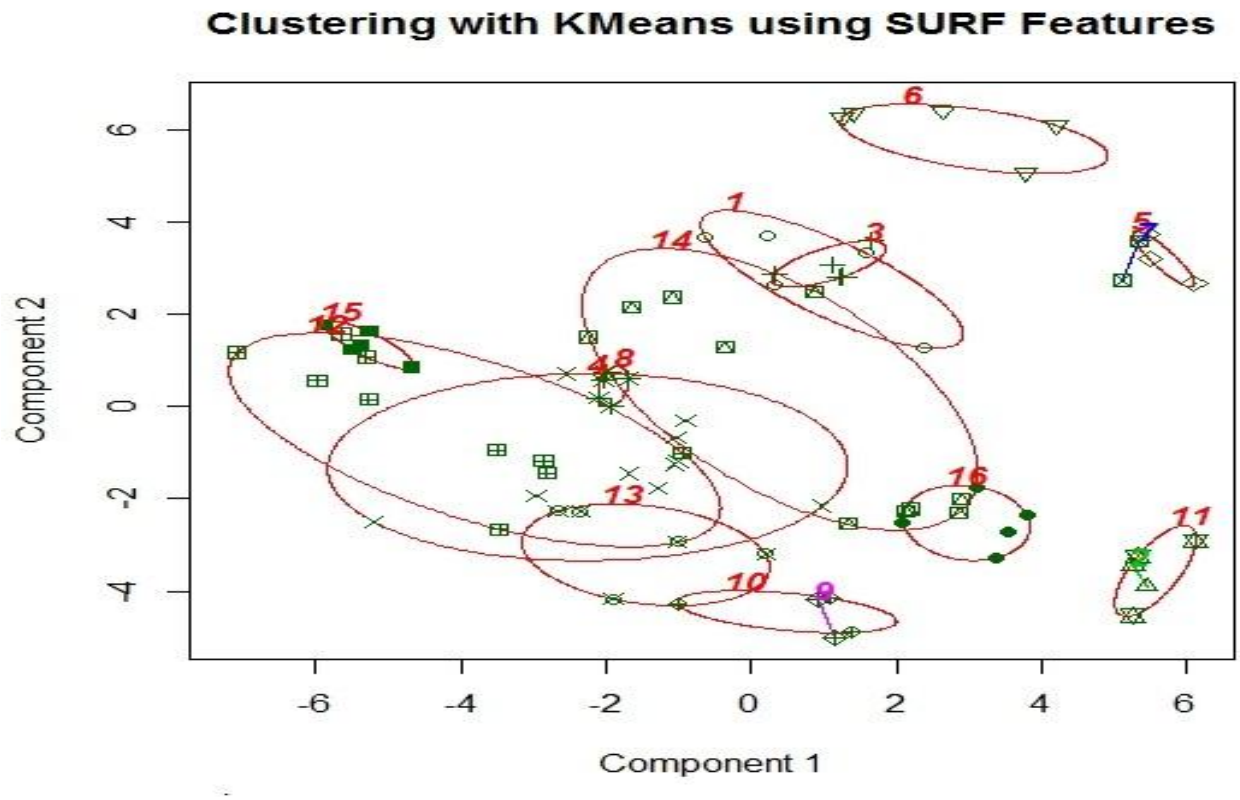
Appendix1: Clustering SIFT features with K Means method for sample characters (14 clusters)



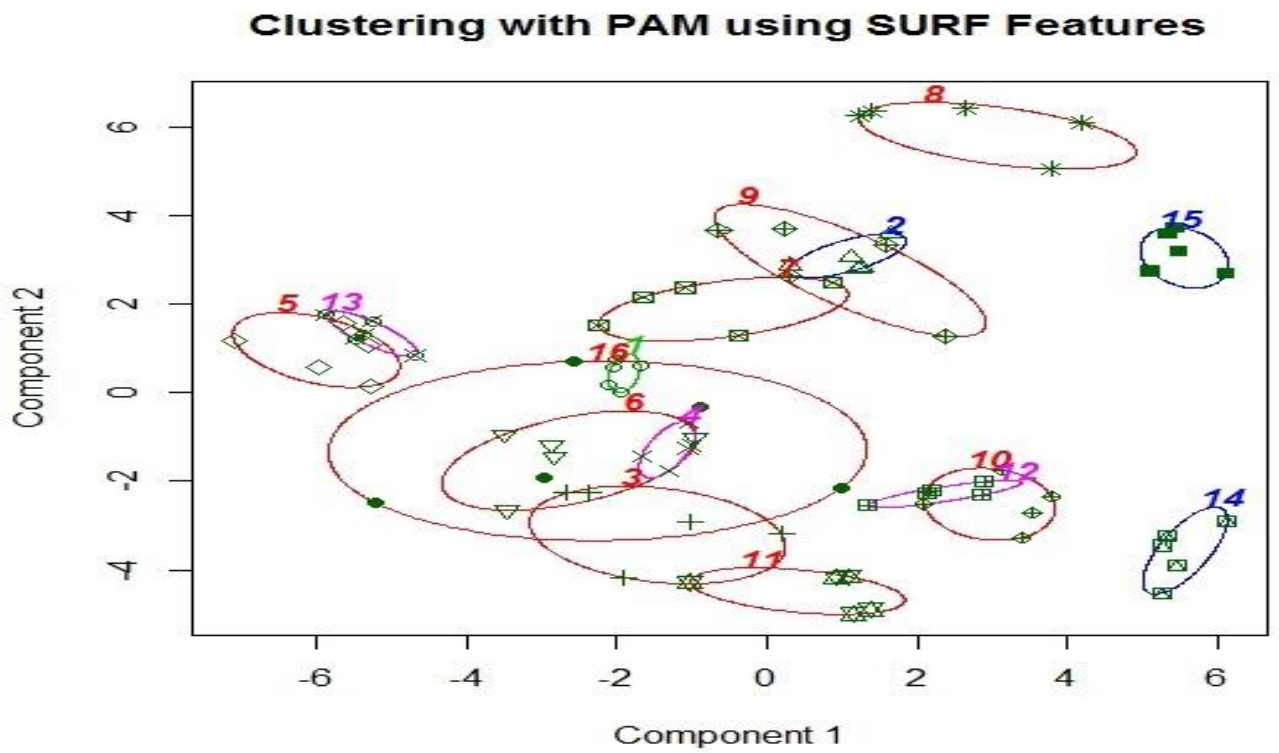
Appendix2: Clustering SIFT features with PAM (Partition around Medoids) for sample characters (14 clusters)



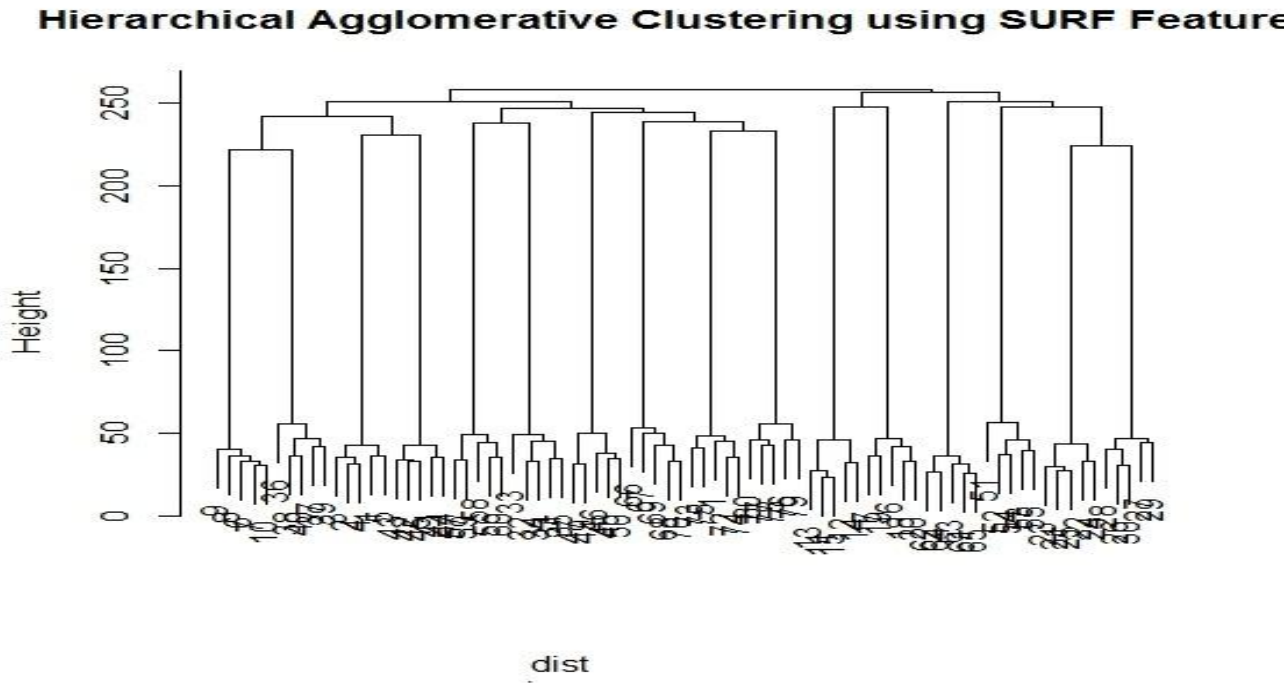
Appendix 3: Clustering SIFT features with Agglomerative Hierarchical clusters for sample characters (16 clusters)



Appendix4: Clustering SURF features with K Means method for sample characters (16 clusters)



Appendix 5: Clustering SURF features with PAM (Partition around Medoids) for sample characters (16 clusters)



Appendix 6:. Clustering SURF features with Agglomerative Hierarchical clusters for sample characters (16 clusters)