

Malayalam Character Recognition using Singular Value Decomposition

Anil R

Centre of excellence in computational engineering and networking, Amrita Vishwa Vidyapeetham, Coimbatore

Arjun Pradeep

Centre of excellence in computational engineering and networking, Amrita Vishwa Vidyapeetham, Coimbatore

Midhun E M

Centre of excellence in computational engineering and networking, Amrita Vishwa Vidyapeetham, Coimbatore

Manjusha K

Centre of excellence in computational engineering and networking, Amrita Vishwa Vidyapeetham, Coimbatore

ABSTRACT

This paper provides a classification methodology of Malayalam characters segmented from scanned document images. Optical Character Recognition (OCR) is one of the successful area which has wide variety of applications related to pattern recognition. This paper describes segmented character recognition using Singular Value Decomposition (SVD). Euclidean distance measure is used for finding the nearest character class of the segmented character image during testing. For each character class, a resultant template is created from training character images using the proposed approach, which in turn reduces the comparisons required for classification. The result obtained from the experiment shows that this method provides an accuracy of 97%.

General Terms

Optical Character Recognition (OCR), Singular Value Decomposition (SVD), Character Classification.

Keywords

Optical Character Recognition (OCR), Singular Value Decomposition (SVD), Malayalam script recognition.

1. INTRODUCTION

Optical Character Recognition is a field of interest for more than a decade and is an active research in pattern recognition [1]. A large number of algorithms have been proposed for the same domain. What it matters is how efficient the algorithm is and how feasible it is. The common pattern recognition technologies are Intelligent Character Recognition (ICR), Optical Character Recognition (OCR), Optical Mark Recognition (OMR), Intelligent Word Recognition (IWR) etc, but among these OCR is the most prevalent technology. In recent years, OCR technology has been applied throughout the entire spectrum of industries, revolutionizing the document management process. The use of OCR varies according to different fields. In banks OCR is used to process checks without the involvement of human. In Health care, to process large volumes of forms for each patient, including their health form and insurance form etc. OCR is also used in many other fields like education, finance, government agencies etc.

Scripts of Indian languages have alphabetic-syllabic nature. The recognition of Indian language scripts are difficult due to large character set, similarity among character classes and existence of both old and new version of language scripts [2] [3]. Malayalam is one of the major languages that come under the Dravidian family of languages. Malayalam characters contain 51 letters and 12 vowel signs which include a lot of similar shaped characters. So classifications among these similar characters always have a room for improvement in accuracy. This paper proposes regional character recognition

of Malayalam characters based on SVD. A new interpretation of SVD based on orthogonality and projection coefficients is used. This paper is organized as follows. Section II gives an overview about Optical Character Recognition. In Section III description about Singular Value Decomposition is presented. Section IV depicts the proposed methodology used for Malayalam character recognition.

2. OPTICAL CHARACTER RECOGNITION (OCR)

Optical Character Recognition (OCR) is the conversion of scanned images of text into machine-encoded text. That scanned images may be handwritten or printed. This is used as a form of data entry from some documents like sales receipts, mail, or any number of printed records. It is a common method of digitizing printed texts so that they can be electronically searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision [4].

Nayana, a product of Centre for Development of Advanced Computing(C-DAC) is an optical character recognition system for printed Malayalam documents. There are some misclassifications because of close-matching characters. For improvement in the existing OCR system a specialized pattern classifier for close-matching characters of Malayalam has been developed [2]. In this, projection histogram combined with Discrete Fourier Transform (DFT) coefficients of close-matching characters are taken as the new feature vector and used for better classification [5].

3. SINGULAR VALUE DECOMPOSITION

In linear algebra, singular value decomposition is a factorization of a matrix to a matrix of lower dimensions with many useful applications. In general SVD can be viewed in three ways. (1) Transforming correlated variables to a set of uncorrelated ones. (2) Identifying and ordering the dimensions along which set of data points exhibit most variation. (3) Factorizing the matrix to a fewer dimension with which we can get a best approximation of the original data points.

An arbitrary real matrix of size $m \times n$ can be decomposed into product of three matrices, U , Σ and V such that

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$$

Where columns of U are Eigen vectors of AA^T , columns of V are Eigen vectors of $A^T A$. The diagonal matrix Σ has

positive Eigen values $\sigma_1, \sigma_2, \dots, \sigma_r$, from AA^T and $A^T A$. These singular values are the square root of nonzero Eigen values. These values are the first r entries on the main diagonal of Σ when A has rank r [6] [7] [8]. SVD is used because approximation of original data points will increase the accuracy of character classification. After SVD, the σ ordered in such a way that with very few values of σ and corresponding columns in U and V better approximates the train image features.

In our experiments, train image size is of 1024, after vectorization. The number of singular values to be taken for optimal error can found using trial and error approach. We have fixed the number of singular values as 100 in all our experiments [9].

4. PROPOSED METHODOLOGY

Scanned Malayalam characters are collected. Each of these scanned images is of size 32×32 . Then each image is vectorized and a matrix is created such that each column in the matrix corresponds to each image vector. This matrix is the database matrix we used in this methodology. It's a herculean task to compare the segmented images from the input to this huge database matrix. Ie, if there are 'n' number of images, then 'n' comparisons will be there. To reduce the number of comparisons mean of 50 different fonts of each character is calculated. A sample image is shown in figure (1) and figure (2) shows how the same image is stored after taking mean [10] [11].

In practical OCR applications, the document to convert is the test image and will be containing sequences of words or paragraphs. The first task here is to segment the image, based on number of lines followed by words and then characters. Using Level set based segmentation approach the test image is segmented into lines, then each line to words and each word to characters [12] [13].

In order to reduce matrix dimension we applied SVD on to the data base matrix. SVD orders the features of the matrix in the order of maximum variation. So with very few values of Σ the whole matrix can be represented. So take on k few values in Σ and create the ΣV^T matrix, which acts as the feature descriptors. For a new character images to determine the character class of the image, the image vector is projected to the k columns of U and the resultant feature is compared with the stored train feature descriptors. Figure (3) shows the layout of the proposed classification.

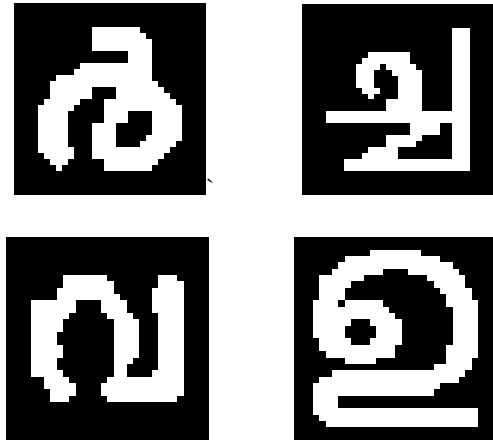


Figure 1: Sample Character Images



Figure 2: Images after taking mean.

4.1 Algorithm

1. Create data matrix, say 'A'
2. Apply SVD, $A = U \Sigma V^T$
3. Compute feature descriptor matrix, $H = \Sigma V^T$
4. Given a test vector 'a' and compute $a^T U$
5. Repeat $a^T U$ to match the dimension of $H = \Sigma V^T$
6. Find the minimum of

$$\left\| \text{repmat}(a^T U, \text{size}(H, 2)) - H \right\|_2$$

to get the desired class.

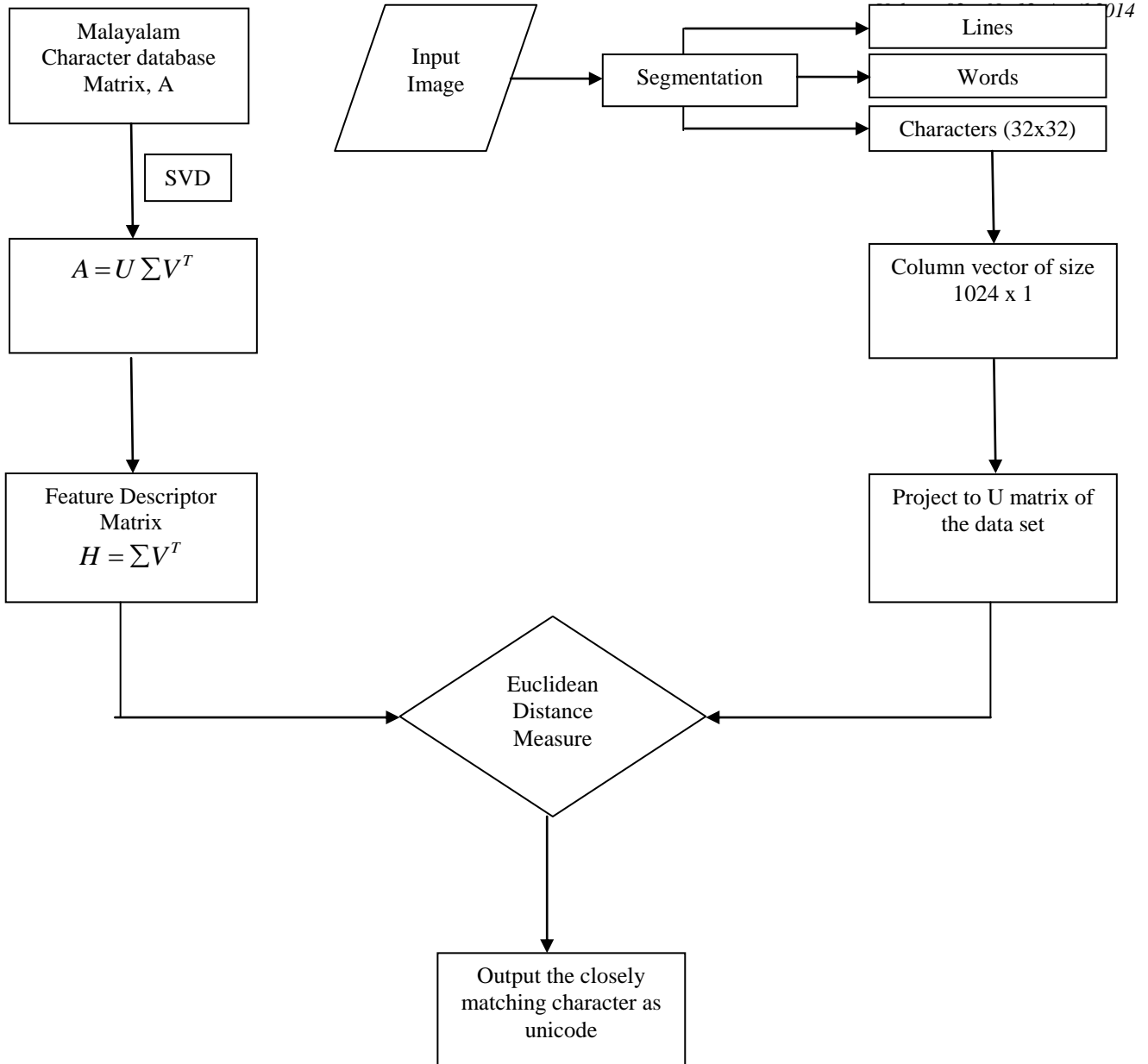


Figure 3: Layout of proposed methodology

5. EXPERIMENTS AND RESULTS

Based on the Malayalam character database, image segmentation combined with Singular Value Decomposition and Euclidean distance is applied to recognize characters in the input image. In this method active contour based segmentation method is used for segmenting the input image [12] [13]. Each lines, words and characters of the input image is segmented separately. Then each segmented characters is tested with the dataset using the proposed approach. Figure (4) shows a sample test image that is used for recognition.

In figure (4), there are 189 characters. From that, 183 characters were classified correctly and rests 6 were misclassified. An accuracy of 96.8 % is obtained. Here the number of comparison is the number of images in the dataset. After taking the mean the number of comparisons is reduced to number of classes in the data set. Here the accuracy is calculated manually but it is less compared to the result in which the mean is not taken. The accuracy while taking mean is 93%. So while taking the mean the accuracy is reduced along

with the number of comparisons. If mean is not taken, accuracy is maintained but number of comparisons will increase.

Here we tested the accuracy with 10 different images. Each image is having a minimum of 180 characters and a maximum of 1000 character. On average for different set of test images, this methodology provides an accuracy of 97%. Table 1 shows the number of characters in the different test images with corresponding accuracy with and without taking mean of each character. Common misclassifications while taking the mean of 50 different font of each character are shown in table 2. The reason for misclassifications is mainly due to the similarity between the shapes of characters. And these misclassifications can be avoided if the mean of each character is not taken.

A dictionary of Malayalam words is created which is used during the classification process. During this process if the word is not present in the dictionary, it will be counted as misclassification.



വളർച്ചയുടെ ഘട്ടങ്ങളിൽ കുട്ടി കൃത്യമായി കാര്യങ്ങൾ ചെയ്യുന്നുണ്ടോ എന്ന് മാതാപിതാക്കൾ ശ്രദ്ധിക്കണം. മൂന്ന് മാസമെത്തിയ കുട്ടി നമ്മെ നോക്കി ചിരിക്കുന്നുണ്ടോ, കൃത്യസമയത്തുതന്നെ ഇരിക്കുകയും മുട്ടിലിഴുകുകയും പിച്ച്വയ്ക്കുകയും ചെയ്യുന്നുണ്ടോ എന്നിവയെല്ലാം പ്രാധാന്യം അർഹിക്കുന്ന കാര്യങ്ങളാണ്.

Figure 4: Sample test image.

Table 1: Numbers of characters in the different test images with corresponding accuracy with and without taking mean

Number of characters in the image	Accuracy without taking mean	Accuracy after taking mean
200	97.8%	95.7%
300	97.3%	94.9%
500	96.8%	94.2%
800	96.3%	94%
1000	96%	94%

Table 2: Common missclassifications while taking mean.

Tested Characters	Misclassified one
	

6. CONCLUSION

Optical Character Recognition is an active area for research and development. This paper focuses on two main tasks: one is segmentation of characters from images and the other is classification of those segmented characters. To reduce the number of comparisons the test image is compared with the mean of different fonts of each characters. To improve recognition performance Singular Value Decomposition (SVD) along with Euclidean distance is used for recognition. The results shows a better recognition performance when the mean of each characters is not taken. When the mean is taken the accuracy along with the number of comparisons for classification is reduced. The classification accuracy of proposed method is 97% when mean is not taken and 94% when mean is taken. To improve the computational performance parallel processing technique can be used.

7. REFERENCES

- [1] N.V. Neeba, Anoop Namboodiri, C.V. Jawahar, P.J. Narayanan, "Recognition of Malayalam Documents". In: *Advances in Pattern Recognition, Guide to OCR for Indic Scripts*, Springer, London, 2009.
- [2] Sajilal Divakaran, "Spectral Analysis of Projection Histogram for Enhancing Close matching character Recognition in Malayalam", *International Journal of Computer Science and Information Technology (IJCSIT)* Vol 4, No 2, April 2012.
- [3] Bidyut B. Chaudhuri," On OCR of a Printed Indian Script. In: *Advances in Pattern Recognition (ed) Digital Document Processing*", Springer London 2007.
- [4] S. Mori, C. Y. Suen and K. Yamamoto," Historical Review of OCR Research and Development[C]", *Proceedings of the IEEE*, 1992, 80(2):1029-1058.
- [5] Stephen V. Rice, George Nagy, Thomas A. Nartker," *Optical Character Recognition: An Illustrated Guide to the Frontier*", Springer, Jan 1999.
- [6] Soumya V. J, Sumya T. Soman, Soman K. P," *Singular Value Decomposition- A Classroom Approach. International Journal of Recent trends in Engineering*", Academy Publishers, Finland, 2009.
- [7] Kirk Baker," *Singular Value Decomposition Tutorial*", March 29, 2005 (Revised January 14, 2013).
- [8] Dan Kalman A Singularly Valuable Decomposition: The SVD of a Matrix. The American University, Washington, 2002.
- [9] Sachin Kumar S, Manjusha K, K. P. Soman, " Novel SVD Based Character Recognition Approach for Malayalam Language Script, *Recent Advances in Intelligent Informatics Advances in Intelligent Systems and Computing* ", Springer, Volume 235, 2014, pp 435-442.
- [10] K. P. Soman, R. Ramanathan," *Digital Signal and Image Processing -The Sparse Way* ", Isa Publishers, 2012.

- [11] Shivsubramani K, R. Loganathan, Srinivasan CJ, Ajay V, K. P. Soman, “ Multiclass Hierarchical SVM for Recognition of Printed Tamil Characters”, In Proc. of IJCAI, 2007.
- [12] Merin Cherian, Radhika G, Shajeesh K U, K. P. Soman, M Sabarimalai Manikandan, ” A Levelset Based Binarization and Segmentation for Scanned Malayalam Document Image Analysis ”, IEEE International Conference on computational Intelligence and Computing Research, 2011.
- [13] Manjusha K, Sachin Kumar S, Jolly Rajendran, K P Soman, “Hindi Character Segmentation in Document Images using Level set Methods and Non-linear Diffusion”, International Journal of Computer Applications, 44(16):42-49, April 2012.