Content Based Image Retrieval System for Malayalam Handwritten Characters

Simily Joseph, Jomy John, Kannan Balakrishnan, Pramod K. Vijayaraghavan
Department of Computer Applications
Cochin University of Science and Technology
Cochin, INDIA
simily.joseph@gmail.com, jomyeldos@gmail.com, mullayilkannan@gmail.com, pramod_k_v@cusat.ac.in

***Abstract*- Content Based Image Retrieval is one of the prominent areas in Computer Vision and Image Processing. Recognition of handwritten characters has been a popular area of research for many years and still remains an open problem. The proposed system uses visual image queries for retrieving similar images from database of Malayalam handwritten characters. Local Binary Pattern (LBP) descriptors of the query images are extracted and those features are compared with the features of the images in database for retrieving desired characters. This system with local binary pattern gives excellent retrieval performance.**

Keywords— CBIR, HCR, Feature Extraction, QBE, LBP Machine Learning, Database

I. INTRODUCTION

Content Based Image Retrieval techniques automate the process of image retrieval in an efficient manner. Conventional text based image retrieval methods exhibits lots of problems. In those systems textual annotations are added manually, it cannot exactly capture the information requirement of user. Text based search gives semantically similar images and content based search gives visually similar images [1]. In recent years the size of multimedia databases has increased rapidly. This leads to the development of better storage and retrieval techniques. The proliferation of internet could meet the need for information to an extent. Often, the unorganized nature of data gives poor performance. In order to meet the requirement of end user development of a Content Based Image Retrieval system is important [2]. CBIR system uses image features for the retrieval of similar images. In the past decades significant progress has been made in the development of CBIR systems. At present CBIR has lot of applications in diverse fields [3]. Some of the main issues in CBIR systems are representation of image, organization of feature vectors, semantic meaning interpretation etc. The mapping from visual feature to perceptual feature is difficult for a machine. [4].

Character Recognition is in the focus of study for decades. It has two branches namely offline recognition and online recognition [5]. Recognition of Indian script is more challenging as it has large character set and high similarity between characters. When dataset is created from real samples of handwritten data, recognition becomes even more complex as large variation is observed in the collected samples. In this paper, we address the problem of retrieval of handwritten Malayalam vowels using content based image retrieval technique. Malayalam is one of the four major Dravidian languages of South India and one among the twenty two scheduled languages of India with official language status in the State of Kerala and Union territories of Lakshadweep and Mahe, spoken by around 30 millions of people and ranked eighth in terms of the number of speakers. Malayalam language script consists of 15 vowels (Fig 1) and 36 consonants

II. RELATED WORKS

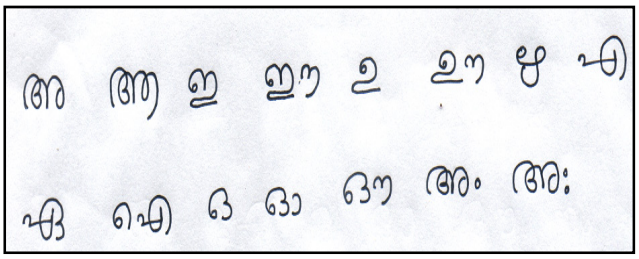

Figure 1: Malayalam vowels

R.C. Veltkamp surveys existing CBIR systems [6]. Amore (Advanced Multimedia Oriented Retrieval Engine) provides the facility of selecting a category of images. Using a kind of template matching the system retrieves similar images. Blobworld allows categorical image search In ImageScape, the user can draw the outline of the desired image. For matching purpose edge mapping methods are used. In iPURE (Perceptual and User friendly Retrieval of Images) initially the images are segmented and then the Individual segments are compared by computing a weighted Euclidean distance. It provides the option of relevance feedback mechanism. MARS (Multimedia Analysis and Retrieval Systems) supports the use of direct queries on low level features. By using queries with Boolean operators, the retrieval accuracy is improved. SQUID (Shape Query Using Image Database) represents the counter of the image using 3 glob shape features. User selects the boundary of an image to retrieve similar image. Several CBIR systems such as IBM's QBIC (Query by Image Content), Virage2, GIFT

978-1-4244-8679-3/11/$26.00 ©2011 IEEE

(GNU image finding tool), IRMA (Image Retrieval for Medical Applications), SPIRS (Spine Pathology and Image Retrieval System) , ImageMap, ASSERT(Automatic Search Research in Character Recognition in Indian Script is still in its early stages. A review of OCR in Indic script can be read in U. Pal [8]. HCR system for Devnagari characters are proposed by S. Arora [9] with a recognition accuracy of about 92.16%. A hybrid zone based feature extraction for recognition of four Indian numerical with nearest neighbor and support vector machine classifiers with a recognition accuracy of 97.85% is reported by S. V. Rajashekararadhya [10]. Another method for recognition of printed and handwritten mixed Kannada numerals is presented using multi-class SVM for recognition yielding a recognition accuracy of 97.76% [11]. Handwritten Tamil Character Recognition sytem using SVM classifiers were proposed by Shanthi [12]. In [13], Fuzzy-zoned normalized vector distance features are classified using class modular neural network considering 44 Malayalam characters. Remarkable works on the application of daubechie wavelet coefficients in HCR was reported by G. Raju [14].

III . PROPOSED SYSTEM

The proposed CBIR system retrieves vowels in Malayalam language. The data sets are collected from different individual without considering age, qualification and profession. Each page is scanned with DPI ranging from 200 to 600 DPI and stored either as BMP, JPG or TIFF format. Each character from the page is separated using morphological [15] operations with rectangular structuring element and the bounding box of each character image is stored as binary image.

and Selection Engine with Retrieval Tools), WebMIRS etc. are also available [7].

A. Noise Removal

Noise is defined as any degradation in the image due to external disturbance. Quality of handwritten documents depends on various factors including quality of paper, aging of documents, quality of pen, color of ink etc. A median filter is used to remove unwanted noise. It is a non linear spatial filter, with output value being the median of the values in the mask. Output of the median filter is a sorted list that have extreme values at the extreme ends of the sorted list. Thus the median filter will replace a noisy value with one closer to its surroundings. Technical details of filtering can be found in[16]. In this paper we have applied median filter with 3x3 mask, to remove almost all unwanted image pixels.

B. Binarization

Binarization is required to concentrate more on the shape of the characters and remove background from the objects. Thresholding is the simplest way of binarization. Given a threshold, T between 0 and 255, replace all the pixels with gray level lower than or equal to T with black (0), the rest with white (1). If the threshold is too low, it may reduce the number of objects and some objects may not be visible. If it is too high, we may include unwanted background information. Otsu's method [17] uses global threshold and the result is found to be satisfactory. So we have used Otsu's method of gray level thresholding.

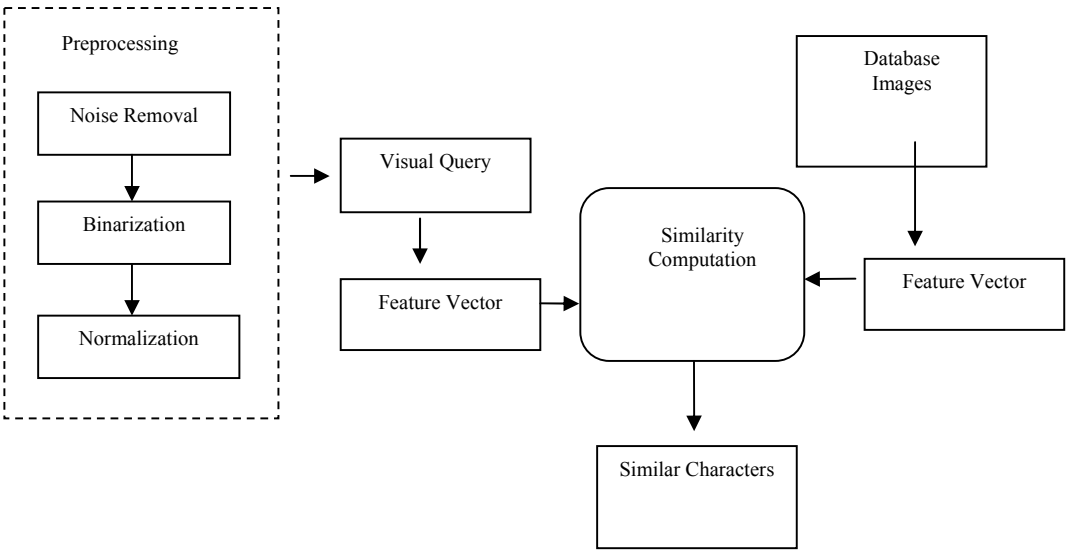

Figure 2: System Architecture

C. Similarity Computation

The application of computer vision to image retrieval has made CBIR a visual information retrieval system using sample image examples. Rather than relying on keywords or any kind of user given metadata, system with Query by Visual Example retrieves images based on the low level features that are extracted from the given images. If the feature description of the query image and database images are similar, the system retrieves those images. The image features are extracted either globally or locally [18,19] . The proposed system extracts Local binary pattern texture descriptors. Texture is the visual patterns of homogeneity [20]. It contains information about the structural arrangements of objects and their relationships. Texture informations can be extracted for individual pixels and for a block of pixels too. In order to reduce the computational complexity, texture extraction methods are usually applied to a block of neighboring pixels. It is concerned with the spatial distribution of gray tones [21].

a. Local Binary Pattern (LBP) Texture Descriptors

LBP (Local Binary Pattern) is introduced as a gray scale invariant to obtain good classification result [22]. The local primitives such as curved edges, points, spot, flat areas etc. can be described using LBP [23]. To generate LBP code for a neighborhood, the weight assigned to each pixels are multiplied with a numerical threshold. The process is repeated for a set of circular samples. As a result the local binary patterns are said to be rotation invariant. Texture over a neighborhood of pixels can be defined as the joint distribution
of the gray value of a central pixel of the neighborhood say g_c and gray value of circular pixels located at distance P.

$$T = t(g_c, g_0, g_{1,\cdots,}g_{p-1} \quad (1)$$

The local texture pattern of a neighborhood can be obtained from the difference of central pixels and each pixel in the neighborhood. As the differences are independent this joint distribution can be factorized:

$$T \approx t(g_c)t(g_0 - g_c,\cdots, g_{p-1} - g_c) \quad (2)$$

To make this invariant against all transformations the signs of the difference are also considered and the overall luminance $t(g_c)$ is ignored as it does not contribute anything to texture analysis.

$$T \approx t(s(g_0 - g_c),\cdots, s(g_{p-1} - g_c)) \quad (3)$$

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

By assigning weight , this difference is converted to a Local Binary Pattern Code which is equivalent to the local texture.

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{p-1} s(g_p - g_c)\, 2^P.$$

This equation results in the generation of 2^PLBP values.

b. Distance Function

The similarity between the query image and the images in the database is obtained by calculating the distance of each feature of the query image and database image. In general, the distance function of query image and database image can be written as $D(Q,P_i)$. Where Q is the query image, P_j is an image in the database. The distance measure used in this study is Euclidean distance, which can be represented using the equation:

$$\sqrt{\sum_{i=1}^{n}((fi(Q) - fi(Pj))^2 * wi)}$$

Where($fi(Q)$ is the i[th] feature of query image , $fi(Pj)$ is the corresponding feature of the image P_j in the database [28].

IV. Result Analysis

In this study query images are selected at random from the test database. Precision is used for evaluating the performance of the system.

$$\text{Precision} = \left| (A \cap B) \right| / \left| A \right|$$

Where A is the set of retrieved images and B is the number of relevant images. The following figure shows the sample results retrieved for different characters. Table 1 shows the average precision obtained for different data set. The performance of this system can be improved by applying relevance feedback mechanism.

Figure 3: Sample results obtained for different query images

Table 1: Precision

Input	Precision				
അ	75%	78%	80%	82%	83%
ആ	78%	80%	83%	88%	100%
ഇ	75%	78%	80%	89%	90%
ഉ	73%	75%	80%	83%	89%
ഊ	77%	78%	83%	90%	100%
ഏ	75%	77%	88%	89%	90%
ഛ	75%	79%	82%	84%	86%
ഒ	80%	83%	86%	93%	100%

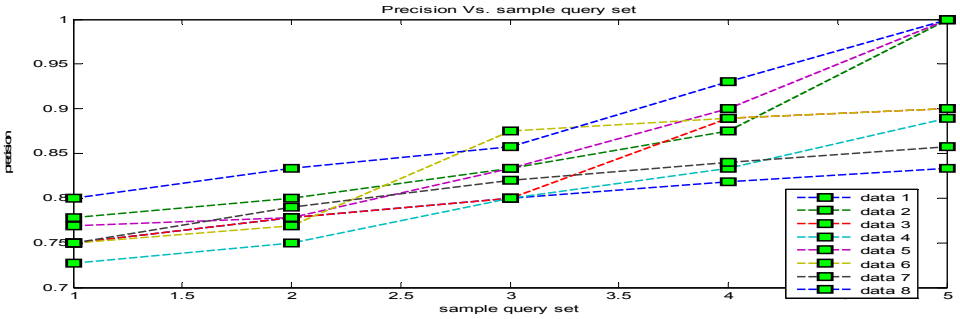

Figure 4 : Plot of Precision Vs Sample Query Set

V. CONCLUSION AND FUTURE WORK

The proposed system is used for the retrieval of similar Malayalam vowel characters. Local Binary Pattern descriptors are used for finding matching characters. The retrieval accuracy can be improved by adding user feedback. The future plan includes use of all characters in Malayalam language. The system can be further enhanced for character recognition.

REFERENCES

[1] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications - Clinical benefits and future directions," International Journal of Medical Informatics,73(1):1–23, 2004.

[2] W. C. Seng , S. H. Mirisaee, Evaluation of a content-based retrieval system for blood cell images with automated methods , J. Med Syst, Springer 2009 ,DOI DOI 10.1007/s10916-009-9393-3.

[3] R .da. S. Torres, A.X. Falcão," Content-based image retrieval: Theory and applications", RITA, Volume XIII, Number 2, 2006, pp. 165-189.

[4] X.S. Zhou, S.Zillner, M. Moeller, M. Sintek, Semantics and CBIR: A Medical Imaging Perspective ACM 978-1-60558-070-8/08/07 ,CIVR'08, Canada.

[5] R. Plamondan, S.N. Srihari, "Online and Offline Handwriting Recognition: A comprehensive Survey", IEEE Trans. On PAMI, Vol22(1) pp 63 – 84, 2000

[6] R.C. Veltkamp, M.Tanase, "Content- based image retrieval systems: A Survey", Technical Report UU-CS-2000-34, October 2000.

[7] P. Aggarwal, H.K. Sardana, G.Jindal, "Content based medical image retrieval: Theory, Gaps and Future", ICGST-GVIP Journal, Volume 9, Issue II, April 2009, pp. 27-37

[8] U. Pal and B.B. Chaudhuri, "Indian script character recognition: A Survey" ,Pattern Recognition, Elsevier ,Vol. 37, pp. 1887-1899, 2004.

[9] S. Arora, D. Bhattacharjee, M. Nasipuri, D K. Basu and M. Kundu,"Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition " , 2008 IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA December 8-10.

[10] S.V. Rajashekararadhya, Vanaja Ranjan P, "Zone-based hydrid feature extraction algorithm for handwritten numeral recognition of four indian scripts",Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics, San Anonio, TX, USA- October 2009

[11] G. G. Rajput, Rajeswari Horakeri, Sidramappa Chandrakant, "Printed and Handwritten Mixed Kannada Numerals Recognition Using SVM", International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010, 1622-1626

[12]N. Shanthi and K. Duraiswamy,"A novel SVM-based handwritten Tamil character recognition system", Theoretical Advances, Pattern Analysis Application (2010) 13:173–180

[13] Lajish V. L., "Handwritten Character Recognition Using Perpetual Fuzzy Zoning and Class modular Neural Networks", Proc. 4th Int. National conf. on Innovations in IT, 2007, 188 – 192.

[14] G. Raju, "Wavelet Transform and Projection Profiles in Handwritten Character Recognition- A Performance Analysis", Proc. Of 16th International Conference on Advanced Computing and Communications, Chennai 2008, pp309-314.

[15] van den Boomgard, R, and R. van Balen, "Methods for Fast Morphological Image Transforms Using Bitmapped Images," Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing, Vol. 54, Number 3, pp. 254-258, May 1992.

[16] Lim, Jae S.," *Two-Dimensional Signal and Image Processing"*, Englewood Cliffs, NJ, Prentice Hall, 1990, pp. 469-476.

[17] Otsu.N, "A threshold selection method from graylevelhistograms", IEEE Trans. Systems, Man and Cybernetics, vol.9, pp.62-66, 1979.

[18] T. M. Deserno, S. Antani, L. Rodney Long, "Content-based image retrieval for scientific literature access", Methods Inf Med 4/2009, pp. 372-380.

[19] T.M.Deserno, M.O.Glld, B.Plodowski, K.Spitzer,B.B. Wein, H.Schubert, H.Ney, T.Seidl.," Extended query refinement for medical image retrieval", Journal of Digital Imaging, Vol 21, No 3 (Sept), 2008:pp. 280-289.

[20] C. H. Chen, L. F. Pau and P. S. P. Wang"The Handbook of Pattern Recognition and Computer Vision" (2nd Edition) , World Scientific Publishing Co., 1998,pp. 207-248.

[21] R.M. Haralick, K. Shanmugham, I. Dinstein, " Textural features for image classification", IEEE Transactions on Systems Man and Cybernetics, Vol.SMC-3, No-6,1973, pp. 610-621.

[22] Harwood D, Ojala T, Pietik¨ainen M, Kelman S & Davis S (1993) Texture classification by center-symmetric auto-correlation, using Kullback discrimination of distributions. Technical report, Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, Maryland. CAR-TR-678.

[23] T. Ojala, M. Pietikäinen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on Pattern Analysis and Machine " Intelligence, vol. 24, no. 7, pp. 971-987, July, 2002