International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 3 Issue 4, April 2016

ISSN: 2394-4404

Optimizing Kannada Content Based Image Retrieval Using Genetic Algorithm

Prof. Nithya E

Assoc. Professor, Dept of CSE,
Dr. Ambedkar Institute of Technology

Kavya G

4th Sem, M.Tech, Dept of CSE,
Dr. Ambedkar Institute of Technology

Abstract: ***Content based image retrieval is the system of searching and retrieving images from a huge dataset. As there are many images in dataset so it is difficult to search for the images that are similar to our query image. Content-based image retrieval (CBIR) approach is required to effectively and efficiently use information from these image repositories. This system helps the users to retrieve the relevant images from the database based on the content. In this paper the data is pre-processed and then we extract Gabor and shape based features. Genetic Algorithm has been used to match the query image features given by users and to recognize the required word in a document.***

Keywords: CBIR; Gabor Features; Genetic Algorithm; Shape based Features; Wiener Filter.

I. INTRODUCTION

Advancement in technology helps us store data in computer based memory and also to retrieve the images from huge database. In an attempt to move in the direction of the paperless office, large quantities of printed files are scanned, digitized and saved as images in databases. In recent years, there has been so much interest in the research area of document image retrieval. The ever increasing amount of multimedia data creates a necessity for new refined ways to retrieve the information one is looking for.

Content based image Retrieval (CBIR) makes use of image own information, most commonly by means of the similarity of image features such as color, texture, shape and structure of the layout, and so on, and the related features of each image for retrieval. Identification of the words in a document image is of primary importance for retrieval. Document image Retrieval systems are available for printed Roman, eastern, Korean, Chinese language, English and different oriental scripts. However, the provision of such products for Indian scripts is still a rarity. The present work addresses the problems involved in designing a font and dimension independent document image retrieval system for printed Kannada textual content.

Kannada is the reputable language of the south Indian State- Karnataka. Realization of Kannada characters is more complicated than many other Indian scripts as a result of higher similarity in character shapes, a higher set of characters and greater variability across fonts within the characters belonging to the same class. The rapid progress of technology has overcome the disadvantage of storing data of physical written information. We have huge data storage systems. Still, a huge number of people would be required to interact more frequently with computer systems. To make the human–computer interaction more potent in such instances, it's desirable to have programs capable of handling inputs in a form of varieties comparable to printed/handwritten paper documents. If the computers have to efficaciously process the scanned images of printed documents, the systems ought to be more sophisticated.

Even though computer systems are used widely in nearly all the fields, surely paper documents occupy an awfully predominant place for a longer period. Additionally, significant share of all sorts of trade writing conversation exist in physical type for various functions, for example, to fax a document, to produce a document in the court, and so on. Therefore, application to mechanically extract, analyze and store expertise from the previous paper form is very much needed for

ISSN: 2394-4404

maintenance and access wherever essential. All these methods go under the title of file image evaluation, which has acquired value as a major research concern in the recent days.

II. RELATED WORK

IWe have studied many previous works done in this field by different researchers. There are many approaches that were followed by different researchers like font and size based, segmentation and binarization based, Characterization of the Image, Entropy based and Correlation based algorithms. Some of the previous works are given below.

Nithya. E. et al (2013) proposed an approach for content based Kannada document image retrieval [1]. Initially they have pre-processed an image to remove the noise present in the input image followed by segmentation. Morphological operation is considered as repeated dilations of an image, called marker image, until the contour of the marker image fits under a second image, called the mask image. Gabor and shape based features are extracted from the training images to match features of query image. In this paper [2] Namrata Dave described few methodologies to segment a text document image. To achieve segmentation of a text based image depends commonly on the presence of guidelines in the document. Appearance of guidelines eliminates the possibility of skew. More over guidelines restricts the character size accordingly of which the total process of segmentation becomes plain sailing [8][9][10]. Mostafizur Rahman and Muzameel Ahmed have together proposed 2-D shape characterization based approach [3] for content based image retrieval. Based on the 3 factors they have retrieved images, first one is morphological operations such as dilation, erosion, opening and closing, second one is to identify the type of images using contour based approach and the third one combines both first and second factors to find similar images in a database. SVM was used as a classifier. Jomy John, Pramod K. V and Kannan Balakrishnan studied on Handwritten Recognition System of South Indian Languages. In this paper [4] they have reviewed multiple papers and algorithms used by each of papers and their results. In this paper [5] Thanuja C and Shreedevi G R proposed Content Based Image Retrieval System for Kannada Query Image from Multilingual Document Image Collection; here they presented visual clues based approach to identify the Kannada text in other languages such as Hindi, English and Maleyalam documents. M.C. Padma and P.A.Vijaya extracted Texture based features for automatic script identification [6].

III. PROPOSED SYSTEM

Fig.1 shows the architecture of proposed system. The effectiveness of CBIR Systems depends on the performance of algorithm for extracting the features of image.

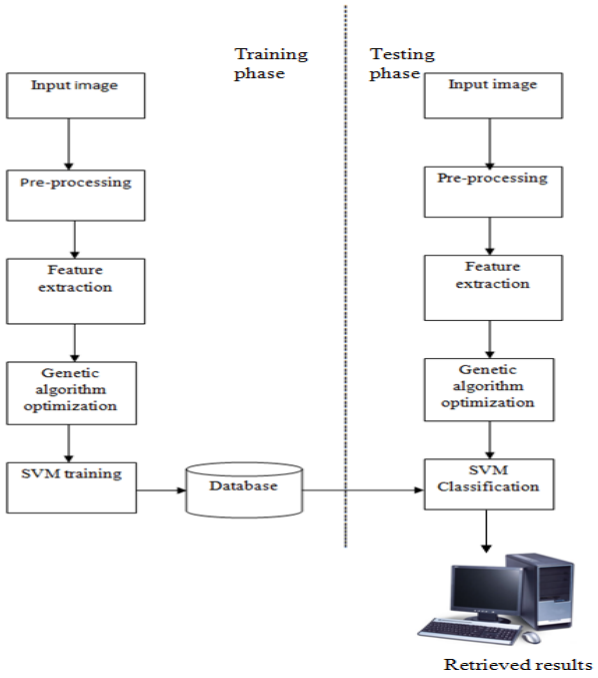

Figure 1: Proposed System

In our proposed work, we divide the work in two phases. i.e. testing and training phase. In the training phase we train the text document by extracting features from the pre-processed images. Extracted features are stored in the knowledge base. In the testing phase, input document is pre-processing includes resize, gray-scale, CLAHE, Wiener filter. Shape and Gabor features extracted are matched stored features using Genetic Algorithm (GA) and retrieve the results. The work can be explained in stepwise as below

A. PRE-PROCESSING

Pre-processing is a stage where we convert the input image to gray scale image, gray values range between 0-255. Contrast of an image is distributed using CLAHE. The noise present in the image can be removed using wiener filter.

CLAHE: In contrast limited histogram equalization (CLHE), the histogram is cut at some threshold and then equalization is applied. Contrast limited adaptive histogram equalization (CLAHE) is an adaptive contrast histogram equalization approach [11], where the contrast of an image is enhanced using CLHE on small data areas referred to as tiles rather than the complete image. The resulting neighboring tiles are then stitched back seamlessly utilizing bilinear interpolation. The contrast in the homogeneous region may also be limited so that noise amplification may also be avoided.

WIENER FILTER: The intention of the Wiener filter is to clear out noise that has corrupted a signal. It is based on a statistical approach. Natural filters are designed for a desired frequency response. The Wiener filter approaches filtering from different angle. One is assumed to have knowledge of the spectral properties of the common signal and the noise, and one seeks the LTI filter whose output would come as close to the original signal as possible.

ISSN: 2394-4404

B. FEATURE EXTRACTION

Gabor features and shape based features have been extracted in the proposed work.

Gabor Features: A two-dimensional Gabor filter is a band-pass spatial filter with selectivity to both orientation and spatial frequency, which can be expressed as follows:

$$G(x,y,\theta_k) = G(x,y)\left[\cos(R) - \exp\left(-\frac{\sigma^2}{2}\right)\right] + iG_1(x,y) * \sin(R) \quad \text{eq. (1)}$$

$$G_1(x,y) = \frac{\lambda^2 \exp\left[-\frac{\lambda^4(x^4+y^4)}{2\sigma^2}\right]}{\sigma^2}, \sigma = \pi \quad \text{eq. (2)}$$

$$R = 2\pi[x\cos(\theta_k) + y\sin(\theta_k)], \quad \lambda = \frac{2\pi}{l} \quad \text{eq. (3)}$$

$$\theta_k = \frac{\pi k}{D}, k = 0,1,2,\ldots\ldots,D-1 \quad \text{eq. (4)}$$

Where l is the wave length, θ_k is the oscillation direction, and D is the number of directions.

In order to extract the Gabor features, initially elastic meshes are constructed on the character image and let the center of each mesh be the sampling point. Then the Gabor feature at the sampling point (x_m, y_m) is extracted as,

$$f_{gabor}(x_m,y_m) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)G(x-x_m,y - y_m; l, v_k) \quad \text{eq(5)}$$

Where M, N, is the size of filter image, $f(x,y)$ is the pixel value on each point (x,y). in practice amplitude of $f_{gabor}(x_m,y_m)$ is usually used as a feature.

C. OPTIMIZING USING GENETIC ALGORITHM (GA)

In the design of automatic pattern classifiers, feature selection and extraction are significant in optimizing performance, and strongly have an impact on classifier design. Ideally, the quandary of feature selection and extraction on the one hand and the classifier design on the other must never be considered independently. Yet, for practical considerations, most researchers make the simplifying assumption that the feature selection/extraction stage and the classification stage are independent. However the ideal goal is proper classification and the intermediate step of feature extraction and dimensionality reduction is, in a sense, subservient to that goal and is not a final result in itself. It might be better to couple feature selection/extraction with effective classification strategies. Then this implies some sort of classification selection feedback mechanism to switch or adapt the feature extractor [12].

D. CLASSIFICATION: SVM

SVM classification method initializes the set with the closest pair of points from opposite classes. As soon as the algorithm finds a violating point in the dataset it greedily adds it to the candidate set. It may so happened that addition of the violating point as a Support Vector may be prevented by other candidate Support Vectors already present in the set. We simply prune away all such points from the candidate set. To ensure that the required conditions are satisfied we make repeated passes through the dataset until no violators can be found. We use the quadratic penalty formulation to ensure linear reparability of the data points in the kernel space. The proposed algorithm is to maximize the network life by minimizing the total transmission energy using energy efficient routes to transmit the packet.

IV. EXPECTED RESULTS AND DISCUSSION

This section explains the results of the proposed system.

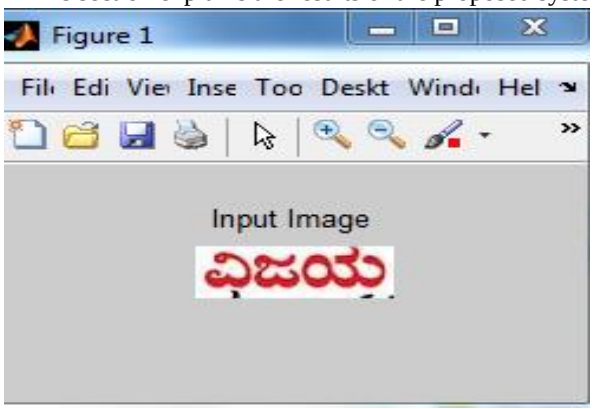

Figure 2: Shows the input image

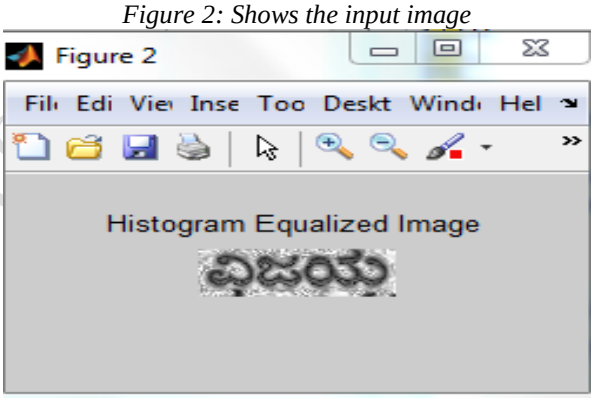

Figure 3: Shows the Histogram Equalization of input image

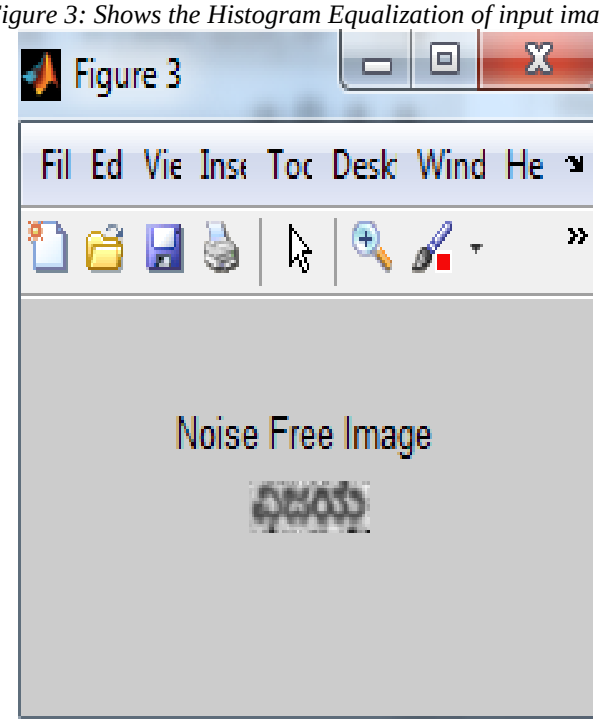

Figure 4: Shows Noise removal of the image

ISSN: 2394-4404

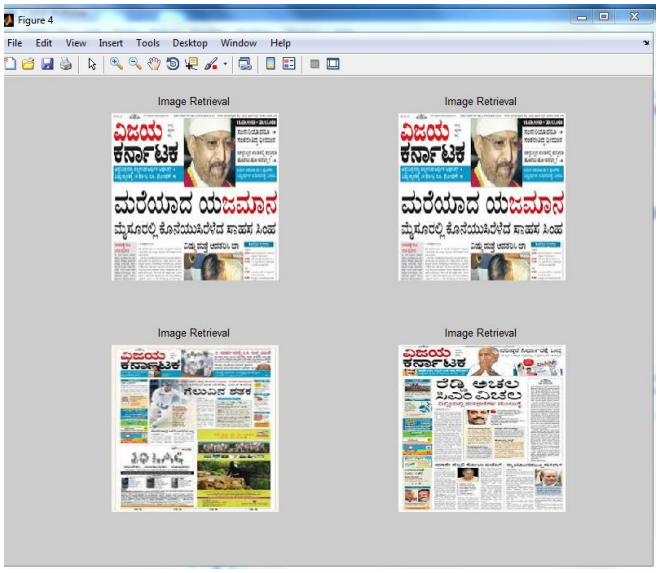

Figure 5: Shows the image retrieved

.

V. CONCLUSION

In this paper we have proposed a new methodology to recognize Kannada characters from document using Genetic Algorithm which is an optimization technique. The computational result is increased by 10% compared to existing system due to the usage of Genetic algorithm. The database size used here is 550KB. Here, a Kannada word accepted in an image format is searched in document consisting of multiple images/pages and then the document page containing these images is retrieved.

REFERENCES

[1] Nithya.E Dr. Ramesh Babu and Chandrakala, "A Font and Size Independent Content Based Retrieval System for Kannada Document Images", Computer Technology & Applications, Volume 4, Issue 2, pp 196-201, 2013.
[2] Namrata Dave, "Segmentation Methods for Hand Written Character Recognition", Volume 8, Issue 4, pp. 155-164, 2015.
[3] Mostafizur Rahman, Muzameel Ahmed, "2D Shape Image Retrieval System Based on the Characterization of the Image", Volume. 4, Issue 8, August 2015.
[4] Jomy John, Pramod K. V and Kannan Balakrishnan, "Handwritten Character Recognition of South Indian Scripts: A Review", 2011.
[5] Thanuja C and Shreedevi G R, "Content Based Image Retrieval System for Kannada Query Image from Multilingual Document Image Collection", Volume 3, Issue 4, pp 1329-1335, Jul-Aug 2013.
[6] M.C. Padma and P.A.Vijaya, "Entropy Based Texture Features Useful for Automatic Script Identification", Volume 2, Issue 2, pp 115-120, 2010.
[7] Chandrakala H T, "A Kannada Document Image Retrieval System based on Correlation Method", Volume 77, Issue 3, September 2013.
[8] Rodolfo P. dos Santos, Gabriela S. Clemente, Tsang Ing Ren and George D.C. Calvalcanti, "Text Line Segmentation Based on Morphology and Histogram Projection", 2009.
[9] Dr.S.Basavaraj Patil, "Neural Network based Bilingual OCR System: Experiment with English and Kannada Bilingual Documents", Volume 13, Issue 8, January 2011.
[10] Ms V. A. Gaikwad and Dr. D.S.Bormane, "An Overview of Character Recognition Focused on Offline Handwriting", Volume 1, Issue 3, December 2008.
[11] Neethu M. Sasi, V. K. Jayasree, Contrast Limited Adaptive Histogram Equalization for Qualitative