Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 11 (2017) pp. 3211-3217

http://www.ripublication.com

Content Based Indian Language Document Image Recovery Using Genetic Algorithm

V.Umesh

Research Scholar, Department of Computer Science,
Bharathiar University, Coimbatore, India

Abstract

Development in numerous technologies and data storage made an increase in the creation of databases. As a result, an appropriate system should be developed to manage these databases. Also, images that are stored in these databases must be retrieved in an efficient way for various applications. The Content Based Image Retrieval (CBIR) system serves this function. In this paper, a system for content based Kannada document image retrieval (CBKDIR) using Genetic Algorithm is introduced. The application of genetic algorithm helps the user in identifying the images that satisfies his needs. Here the data is pre-processed and then Gabor features and shape-based features are extracted. Using this algorithm, an effective search and retrieval of scanned Kannada document images from a huge of collection of databases can be achieved.

Keywords: CBIR, Gabor Features, Genetic Algorithm, Shape based Features.

1. INTRODUCTION

Technology is the speed of light. Every now and then it is getting updated. Today world speaks on digitalization. Here we try to keep anything and everything on web so dealing with messy papers is evaded. So all documents are scanned and kept in knowledge base. Lot of researches is going on CBIR. Multimedia images are becoming enormous. So optimizing this kind of document image retrieval is a vital role.

Content based image Retrieval (CBIR) will handle information of the images and check for the similarities in those information. European scripts are easier in this regard and best approaches such as Optical Character Recognition (OCR) for retrieval are already present. But Indian scripts pose loads of hurdles and obstacles. The present work is to

address the problems involved in designing font and dimension of the document image retrieval system for printed Kannada textual content.

Kannada is the esteemed language of south Indian State in particular it is Karnataka. Comprehensions of Kannada characters are more awkward than many other Indian scripts. And also resemblance in shapes, characters and superior inconsistency across fonts within the characters belonging to the same class adds to the complexity of Indian scripts especially for Kannada language.

Written documents are also set to this kind of storage mode. Many of the commercial bills need to be preserved long term. This approach can be deployed there also. Many people transact in a business so many will access those document data. The interaction of system and the people should be dealt with ease here. It's desirable property of programs to be able to handle inputs in a form of versatile comparable to printed and also handwritten paper documents. If the computers have to process properly the scanned input images of printed documents, the systems must be sophisticated.

Addressing all these problems, this paper proposes an efficient approach for the retrieval of content based Kannada image using genetic algorithm- which is an optimizing technique.

2. LITERATURE SURVEY

Many previous works done by different researchers in this field have been studied. In recent years, a number of different approaches have been proposed for efficient search and retrieval of document images from large database such as font and size independent, correlation based, characterization of the images, segmentation and binarization based. Some of the works regarding this field is given below.

Nithya.E et al (2013) proposed font and size independent content based image retrieval of Kannada document [1]. Primarily they have pre-processed an image to remove the noise present in the input image followed by segmentation. To match the feature of the query image they have extracted the Gabor and shape based features from the training images. Here, they have showed that the segmentation algorithm is robust to the noise image. In this paper [2] Ashour proposed a new content based image retrieval method using color and texture features in order to avoid the problems of computational complexity and the retrieval accuracy. Here they have calculated the color moment and texture moment to extract the color and texture features.

In paper [3] C.Ramesh Babu Durai, V.Duraisamy, C.Vinoth Kumar has proposed an approach for content based retrieval of image using neural network in medical image classification. They collected the images from scan centre and they also designed roulette wheel such that a sector of the wheel is occupied by the population using the fitness value. 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 Malayalam documents. M.C.Padma and P.A.Vijaya extracted Texture based features for automatic script identification [6].

3. EXISTING SYSTEM

The existing system proposed an algorithm for retrieval of document image using phase based image matching – which is a technique for matching the image components of phase in Fast Fourier Transformation by which the phase angle of input image and query image is determined that helps in matching word for the retrieval of document. This system accepts a textual query from users. The textual query is first converted to an image by rendering, features are extracted from these images and then a search is carried out for retrieval of relevant documents. Results of the search are pages from document image collections containing the retrieved words sorted based on their relevance to the query. This work mainly aims at addressing some of the issues involved in effective and efficient retrieval in document images with effective representations of the content Kannada word images.

4. PROPOSED SYSTEM

Progress in technology helped us to store records in database and also to retrieve images from massive folders. In an effort to move in the way of the paperless office, large quantities of printed documents are scanned and saved as images in the databases.

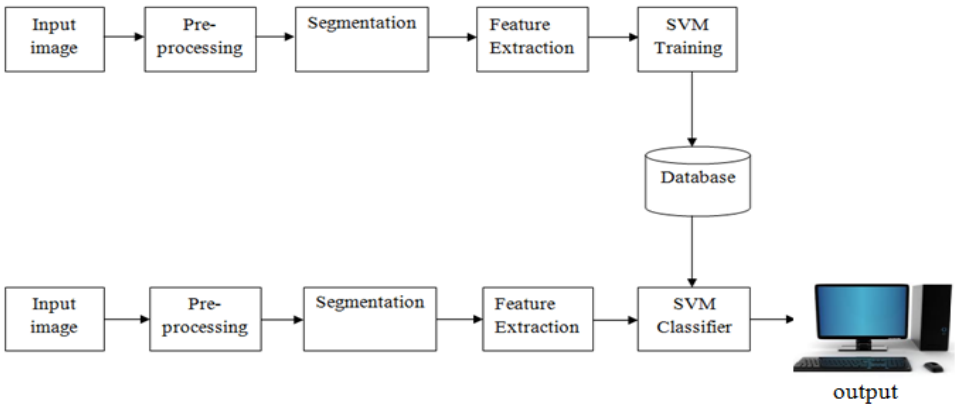

Fig.1. Proposed Architecture

In our proposed work, we are actually dividing the work in two phases. They are, testing and training phase. In the training phase we have to train the text document by discovering the features from the pre-processed digital images. These features are then put into the store of our data. In the testing phase, input document is pre-processed which includes resize, gray-scale, CLAHE, Weiner filter. Shape and Gabor features extracted are matched stored features using Genetic Algorithm (GA) and retrieve the results.

5. IMPLEMENTATION

An efficient system for retrieval of Kannada document images using genetic algorithm is presented in this paper. The work can be explained in stepwise as below:

5.1 PRE-PROCESSING

Pre-processing is a step where we convert the inputted digital image to gray scale image. These gray scale images will have gray values ranging between 0-255. Contrast feature of an image is distributed using CLAHE. The noise present in the image can be removed by simple usage of wiener filter.

CLAHE: In contrast limited histogram equalization (CLHE), the histogram is cut at some conditional point and then step of equalization is applied. Contrast limited adaptive histogram equalization (CLAHE) is an adaptive contrast histogram equalization process [11], where the contrast of an image is tried to enhance using CLHE on small data areas. We call these data areas as tiles instead of full 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 escaped.

Wiener Filter: The purpose of the Wiener filter is to erase noise that has destroyed a signal. It is assumed that signal and (additive) noise are stationary linear random processes. It requires being physically realizable. Its Performance criteria demands minimum mean-square error. Wiener filters are moderately very slower to apply. This is because they work in the frequency domain. To enhance filtering, one can opt the inverse FFT of the Wiener filter G (u,v) to get an impulse response g (n,m). This impulse response can be skipped to yield a convolution mask. The spatially truncated Wiener filter is substandard to the frequency domain version. The blur in mage can be because of linear motion or unfocussed optics. The technique for removal of blurring in digital images is the Wiener filter. From a signal processing standpoint, blurring due to linear motion in a photograph is the result of poor sampling. Each pixel in a digital representation of the photograph should epitomize the intensity of non-moving point in front of the digital camera.

5.2 FEATURE EXTRACTION

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

Gabor Features is filter with selectivity to both direction and longitudinal frequency. It is expressed as below:

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

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

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

$$\theta_k = \frac{\pi k}{D}, k = 0,1,2, \ldots\ldots, D-1 \tag{4}$$

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,

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

Where M, N is the size of filter image and $f(x,y)$ is the pixel value at each point (x,y).

5.3 OPTIMIZING USING GENETIC ALGORITHM (GA)

In the design of programmed pattern classifiers, feature selection and extraction are significant in optimizing performance, and strongly have an impact on classifier design. The dilemma of feature selection and extraction and the classifier design must never be considered independently. So, forth considerations, practically most researchers will make the simplifying assumption about the feature selection and the classification stages are independent. However the superlative goal is proper classification and the midway step of feature extraction and dimensionality reduction is submissive to that goal. It is better to combine feature extraction with effective classification strategies. Then this will imply some sort of mechanism of classification selection feedback in order to switch the feature extractor [12].

5.4 SVM CLASSIFICATION

SVM classification will initialize the collection of closest pair of points within opposite classes. Very sooner the algorithm finds irreverent point in the generated dataset it voraciously adds them to the candidate set. The inclusion of the violating point in terms of a Support Vector may be taken out by other candidate Support Vectors already in the set. We so prune all those points from the set of candidate. To ensure about required conditions we will make repeated passes through the dataset until no violators come on

our way. We use the formulae of quadratic penalty to make sure linear reparability of the data points in the kernel space.

6. EXPERIMENTAL RESULTS

This section explains the results of the proposed system.

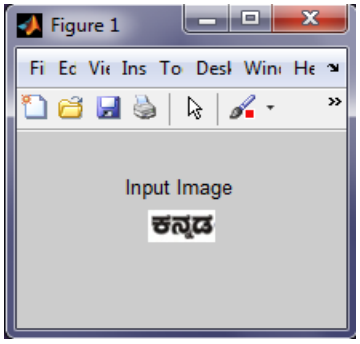

Fig.2. Shows the input image

Fig.3. Shows the Histogram Equalization of input image

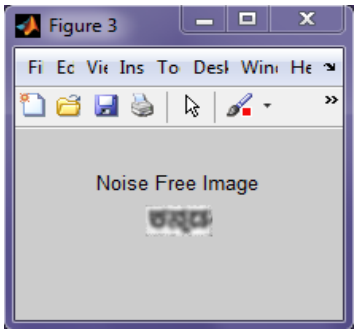

Figure.4. Shows Noise removal of the image

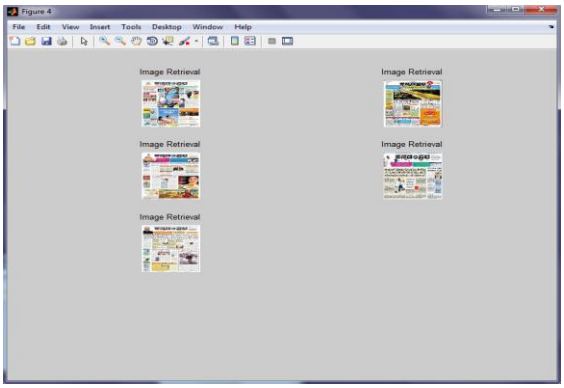

Figure.5. Shows the image retrieved

7. CONCLUSION

In this paper we have presented a systematized method for the retrieval of Kannada document image from a large collection of scanned Kannada document by the use of Genetic algorithm as an optimization technique. The computational result is increased by 10% compared to the existing system due to the usage of Genetic algorithm. This work finally produces a robust and flexible retrieval system.

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] C.Ramesh Babu Durai1, V.Duraisamy2, C.Vinothkumar3, "Improved Content Based Image Retrieval using Neural Network Optimization with Genetic Algorithm", ISSN 2250-2459,Volume 2, Issue 7, July 2012.

[3] R.C.Joshi, and Shashikala Tapaswi, "Image Similarity: A Genetic Algorithm Based Approach", World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol: 1, No: 3, 2007.

[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] Mostafizur Rahman, Muzameel Ahmed, "2D Shape Image Retrieval System Based on the Characterization of the Image", Volume. 4, Issue 8, August 2015.

[7] Namrata Dave, "Segmentation Methods for Hand Written Character Recognition", Volume 8, Issue 4, pp. 155-164, 2015.

[8] M.C.Padma and P.A.Vijaya, "Entropy Based Texture Features Useful for Automatic Script Identification", Volume 2, Issue 2, pp 115-120, 2010.

[9] Chandrakala H T, "A Kannada Document Image Retrieval System based on Correlation Method", Volume 77, Issue 3, September 2013.

[10] 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.

[11] Dr.S.Basavaraj Patil, "Neural Network based Bilingual OCR System: Experiment with English and Kannada Bilingual Documents", Volume 13, Issue 8, January 2011.