# A Soft Computing Technique for Foraging Retrieval of Kannada Word from Scanned Document

Mr. Nithya. E
Assoc. professor, CSE department
Dr. Ambedkar Institute of Technology, Bangalore-560056

Nayana K S
4th Sem, M.Tech, CSE Dept
Dr. Ambedkar Institute of Technology, Bangalore-560056

Dr. Ramesh Babu D R
Prof. & HOD, Dept of CSE-DSCE
Bangalore

***Abstract*— This work presents a new idea for retrieving a Kannada word from a large database of scanned Kannada document. The need for effective Kannada document image retrieval aims at improving the efficiency and accuracy of retrieval by using genetic algorithm. This overcomes the problem of deliberate retrieval and segmentation, especially for Kannada characters. Genetic algorithm includes three stages with some predefined set of initial population. In this paper, retrieval results are the locations at which the query image is present.**

***Index Terms—Document Image Retrieval, CBKDIR, Genetic algorithm, optimization techniques.***

## I. INTRODUCTION

Document Image Retrieval (DIR) is the area of research with high interest. This is due to the introduction of Digitization, which in turn provides a proficient way to process, safeguard and communicate all types of data. Hence there is a need for retrieval of those digitized data. Increase in the amount of stored data creates a need for new optimized methods for fast retrieval of specified information.

Two methods were proposed for Image retrieval: Tag Based and Content Based. Content Based Image retrieval is the most convincing method due to less accuracy in tag-Based, which is highly dependent on tag creator of an image.

The availability of CBIR systems for printed foreign languages like Roman, Japanese, Korean, Chinese, and English exists but these systems are rarely available for Indian scripts especially for South Indian Language Kannada. Kannada is the bureaucrat language of the south Indian State-Karnataka. Since Kannada is a language with high complex characters (vattaksharas) and the characters with similar shapes, searching of those characters in a scanned Kannada document is highly difficult.

Generally CBDIR can be done as shown in the figure below:

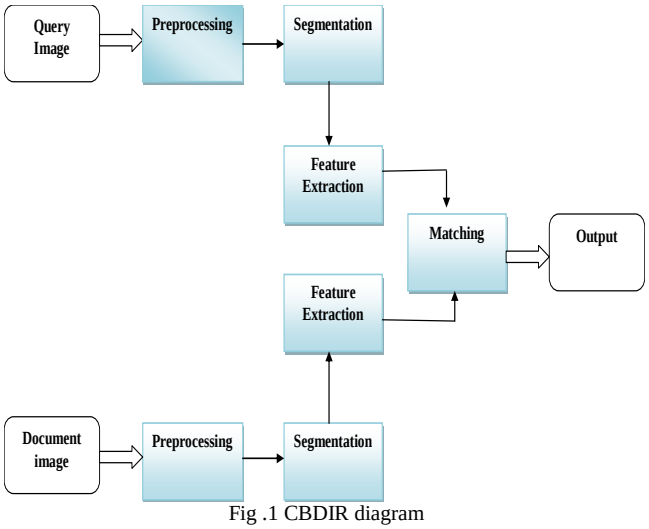


Fig .1 CBDIR diagram

Even though many research works has been done with respect to this Kannada Document Image Retrieval (CBKDIR), none of them have got a firm place due to some drawbacks like slow retrieval, inefficient retrieval and etc.

The present work addresses these drawbacks by the use of some soft Computing techniques as an optimization for Retrieval. One of the soft computing techniques we are using here is Genetic algorithm. Genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a Meta heuristic) is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms fall into class of evolutionary algorithms (EA), which engender solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, selection, crossover and mutation.

In this paper, we propose an effective retrieval method for Kannada document image by the use of Genetic Algorithm as an optimization technique.

## II. RELATED WORK

Image processing was origin from 1960s, at the Jet Propulsion Laboratory and a few other research facilities, and they have enriched their techniques with application to

satellite imagery, medical imaging, character recognition, and photograph enhancement [1].

A new manner of image processing was developed in 2002, termed as GDIP (Gradient Domain Image Processing) which manipulates pixel values and differences between pixels.

G. Rafiee et al. [3] proposed CBIR as a technique in which three correlated modules including path sampling, characterizing and recognizing are used.

Thomas M. Deserno [4] defines ontology of 14 gaps in which the content of the image, image features and the performance of the system while processing and its usability were employed. Using artificial intelligence and mathematical processing, they have proposed that there always exits a semantic gap from the low level computer processing to the human analysis of pixels.

T. Dharani et al. [5] proposes CBIR in a different manner, in which they use visual appearance of image such as color, shape, texture etc., to search query image specified by the user, from large database.

Schaefer G. [6] reviews a number of fundamental CBIR algorithms that gain color, texture and shape features, and they proposed the method for extracting features from images like JPEG formats without the use of decompressing.

The system for the retrieval of kannada text as an image from a database of scanned kannada document was proposed by Nithya E et al. [7]. FFT has been implemented to find the phase angle of the given query image.

Thanuja C et al. [8] have proposed the visual clues based procedure to identify Kannada text from a multilingual document which contains other languages along with Kannada.

John Holland in 1960 [9] is the one who have proposed Genetic algorithms (GA) for the first time. Holland used an efficient algorithm which can automatically search and extract query image from the database of images with complex background.

Mathematical morphology has produced an important class of non-linear filters. It [10] describes research into methods by which different classes of morphological filters can be designed by employing genetic algorithms.

The retrieval of images from large database based on the multi-feature similarity score fusion can be done effectively by using genetic algorithm [11]. Fusing multi-feature similarity score is expected to improve the system's retrieval performance [12].

GAs are proved as the most powerful, unbiased optimization techniques for sampling a large solution space[13], because of unbiased stochastic sampling, they were quickly adopted in Image processing.

Image Optimization and Prediction [14] is the combination of features of Query Optimization, Image Processing and Prediction. In this paper, they have used the GA as a method of optimization.

This paper depicts an optimized CBIR system that uses multiple feature fusion and matching to retrieve images from an image database [15]. Traditional optimization methods can solve a wide array of problems, but it is also this generality that often prevents them from dealing with large data efficiently [16].

## III. EXISTING SYSTEM

This system proposed a document image retrieval algorithm using phase based image matching – an image matching technique using the phase components in Fast Fourier Transformation which determines the phase angle of input image and query image that helps in matching word for the retrieval of document.

The query text is given to the system from users and then it is converted into JPEG format. It renders and extracts the features from that input image and it searches for relevant documents for retrieval.

Here, the search is based on matching the query image with the database images by using phase based method for matching images. It also delivers word from the source image matching the query. This phase has input as two images, source image and query image, which are converted into binary form. First, Fast Fourier transform is performed on both the images and phase angle of both the images are determined.

Then, subtracting the differential phase angle from first image to second image leads to perform Inverse FFT which gives phase difference. If the result of IFFT breaks threshold value then words are matching, else the words are not matching.

In this section, the principle of phase-based image matching using the limited Phase-Only Correlation (POC) function (which is sometimes called the ―phase-correlation function‖) is used.

Fast Fourier Transform (FFT) algorithms have computational complexity O (n log n) which is better than O ($n^2$) and proportionally provides stable results.

Direct matching of pixels in images is inefficient due to the complexity of matching and thus impractical for large databases. This problem is solved by directly storing word image representations.

The effect of combination of different fonts in a single collection can be one possible direction for exploring the feasibility of the proposed technique.

s

## IV. SYSTEM ARCHITECTURE

Since the retrieval speed is considerably less in the existing system, it is not possible to achieve effective results. It also requires more number of initial samples and large solution space. By considering these snags of existing system, we have proposed a system with the implementation of Genetic algorithm which makes the retrieval faster with accuracy.

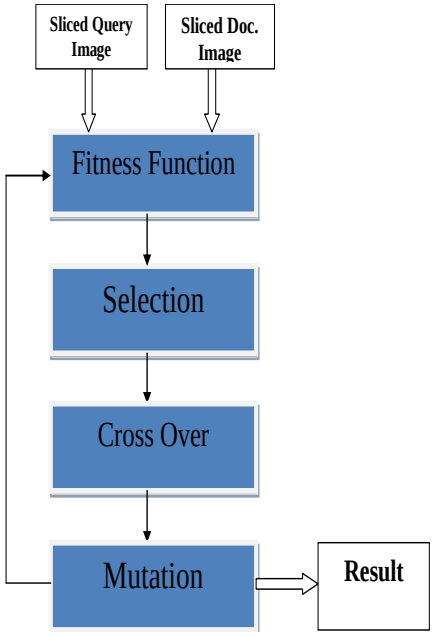


Fig.2 Proposed System architecture

Above figure shows the architecture of the proposed system, which mainly contains three components: Input component, System and an output component. System contains Genetic Algorithm procedure which will be explained in implementation.

## V. IMPLEMENTATION

This section provides information about the phases involved in this work. Initially it contains three components: Input component which provides two inputs to the system, System which is embedded with Genetic algorithm and an output component which shows the location of the query image in the document image.

*Input Component*

Two inputs are given to this proposed system i.e., a query image and a document image. This query image is sliced into number of pieces to extract each feature and then a cropped document image of size query image is also given as input to the system.

Document image is cropped into query image size which results in number of cropped images. This can be done by using "I=imcrop (DImg,[xmin,ymin,width,height]);". After cropping is done, we have sliced each cropped document image into number of slices.

Providing this sliced query image and one of the cropped document images (sliced) into the system as inputs boosts up the system to proceed with next process of retrieval.

*System embedded with Genetic algorithm*

The proposed system mainly concentrates on Genetic algorithm (GA), which is a method for solving both guarded and unguarded optimization problems based on a natural selection process that imitates biological evolution.

Genetic algorithm is an encapsulation of three main processes: *Selection, Crossover and Mutation.*

Initially the GA is provided with some set of randomly generated initial samples of population.

*SELECTION STAGE:*

This is the stage at which the algorithm selects a few number of population based on the fitness value of initial samples.

As soon as this stage gets the input of initial samples, it calculates the fitness value. This is the value obtained by comparing each slice of query image with every slice of cropped document image. The population with higher fitness value will be selected for next stage.

*CROSSOVER:*

It is a process of generating a new kind of population points by taking more than one parent solution.

In this stage, it takes two or more population points from the output of selection stage then it produces a child solution from those parents by using one of the crossover operators. We are using the fitness function as a crossover operator. The newly generated child solution is the combination of its parents' features and hence it is the one with higher fitness than its parents.

After generating new set of population points, it computes the fitness value and the fittest population will be moved to the next stage.

The below example shows the process of crossover by considering ones and zeros as chromosomes (features).

Parent1: 1 0 1 1 0 0 1
Parent2: 0 1 0 1 0 1 0
Parent3: 0 0 1 1 0 0 1

Produces the following child,

Offspring: 0 0 1 1 0 0 1

In the above example, the new offspring is generated from three parents. Initially it checks the features/chromosomes of parent 1 and parent 2, if they are same that will be remained for the offspring otherwise parent 3's feature will be chosen for offspring.

*MUTATION:*

Mutation is a change that happens to maintain genetic diversity from one generation to next generation.

The population set of points generated from cross over is taken as input and the genetic sequence of those population has been changed based on some mutation operator to balance genetic diversity.

The below example shows how the genetic sequence has been changed after mutation:

1 0 0 1 0 0 1 ↓

1 0 0 1 0 1 1

The mutation of bit strings happens at random positions. The probability of mutation of a bit is 1/n, where n is the length of the bit string/ gene sequence.

After obtaining the resultant population from the mutation stage, fitness value will be calculated for those resultants. Then the population with higher fitness will be moved to the next generation. The population selected for next generation again undergone through selection, crossover and mutation. This process continues iteratively until the population "evolves" toward an optimal solution.
Here the location of the query image in the document image is the optimal solution.

## VI. CONCLUSION

This paper proposed a systematized method for the retrieval of Kannada document image from a large database of scanned Kannada document. The computational result is increased by 10 % compared to the existing system due to the usage of Genetic algorithm which is an optimized technique. Since GA takes very less number of initial samples, it requires less solution space and computationally faster. Future work can be done for the usage of histogram for computing fitness while comparing each slices of query and document image.

## REFERENCES

[1] Azriel Rosenfeld, Picture Processing by Computer, New York: Academic Press, 1969
[2] Bhat, Ranan, Pravin, et al. "Gradientshop: A gradient-domain optimization framework for image and video filtering," in ACM Transactions on Graphics (TOG) 29.2 (2010): 10.
[3] Rafiee. G, Dlay. S. S, Woo. W. L., "A review of content-based image retrieval," in Communication Systems Networks and Digital Signal Processing (CSNDSP), 2010.
[4]. Thomas M Deserno, Sameer Antani and Rodney Long, "Ontology of Gaps in Content-Based Image Retrieval" in 2007.
[5]. T.Dharani Aroquiaraj I.L.,"A survey on content based image retrieval" in PatternRecognition, Informatics and Mobile Engineering (PRIME), 2013.
[6] Schaefer.G.,,"An introduction to content-based image retrieval" in Digital Information Management (ICDIM), 2013.
[7] Mr. Nithya. E, Dr. Ramesh Babu D R, "Content Based Kannada Document Image Retrieval (CBKDIR)" in International Journal of Engineering Research & Technology (IJERT), 2013
[8] Thanuja C, Shreedevi G R, " Content Based Image Retrieval System for Kannada Query Image from Multilingual Document Image Collection " in International Journal of Engineering Research and Applications (IJERA) ,2013
[9] John H Holland," Genetic algorithms and adaptation" in Adaptive Control of III-Defined Systems, NATO Conference Series, 1984
[10] N.R.Harvey, S.Marshall,"The design of different classes of morphological filter using genetic algorithms" in Image processing and applications (IEEE), 1995
[11] Mianshu Chen, Changchun, Ping Fu, Yuan Sun, Hui Zhang, " Image retrieval based on multi-feature similarity score fusion using genetic algorithm" in Computer and Automation Engineering (ICCAE), 2010
[12] Sapthagiri.k, Manickam.L, "An Efficient Image Retrieval Based on Color, Texture (GLCM & CCM) features, and Genetic-Algorithm" in International Journal Of Merging Technology And Advanced Research In Computing,20
[13] Mantas Paulinas, Andrius Ušinskas," A Survey Of Genetic Algorithms Applications For Image Enhancement And Segmentation" in Information Technology And Control, 2007
[14] Ms. Sweta V. Jain, Urmila Shrawankar,"Image Optimization and Prediction" in International Conferences CAAM-09, 2009
[15] R.Priya and Dr.Vasantha Kalyani David, "Optimized Content based Image Retrieval System based on Multiple Feature Fusion Algorithm" in International Journal of Computer Applications, 2011.
[16] Katsuhiko Sakaue, Akira Amano and Naokazu Yokoya "Optimization Approaches in Computer Vision and Image Processing" in IEICE Trans .INF. & SYST in 1999