IMAGE MINING: SIMILARITY SEARCH BASED ON GENETIC OPTIMIZATION ALGORITHM FOR LARGE SCALE DATABASE

[1] T.Sowmi, [2] E. Saravana Kumar
[1] PG Scholar, [2] Associate professor
[1] Department of Computer Science and Engineering
[1] Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

Abstract **— In digital technology, the usage of photo sharing websites are increased in recent years. So the multimedia data's which has been communicated by the user are dumbed frequently. From the storage, retrieving the relevant information for the query became the challenging task among the research communities. In prior novel NDH method, the recognition rate was low in terms of detecting the similarity and dissimilarity among the images for the query sample. It also acquired more iterative procedure to obtain the optimized subset by updating the transformation matrix. So to improve the performance, the genetic optimization algorithm is incorporated. The optimized feature subset is obtained by the fitness value by considering the four feature extraction methods. The extensive experiments are carried out with four datasets and compared with the state-of-the-art techniques in MATLAB.**

Keywords **— Genetic Algorithm, NDH, Transformation matrix, optimized subset**

I. INTRODUCTION

Due to the enormous development of digital technologies, the usage of multimedia data's such as video, audio etc., are increased [1]. So retrieving the relevant images became the vital role in multimedia research community. Data mining [2] refers to extracting the needed information from the data warehouse which undergoes four process as problem definition, data gathering and preparation, model building and knowledge deployment. These process refines the data and extracts the needed information.

Image mining [4] is defined as extracting the relevant patterns from the images which are not directly found in the images. Comparing with the text based retrieval techniques, image retrieval [3] process undergoes a major consequence in digital technologies. In earlier days, many techniques related to the image retrieval process were used. Particularly, in hashing based optimization technique, the recognition rate was low interms of precision and recall [24]. Hashing based technique was also considered as an iterative process which undergoes many sequences [1]. For each iteration, the transformation matrix for the whole dataset was updated and also the computational complexity was also high due to these sequences. Locality sensitivity hashing [5] technique was used for the binary bit generation which is indicated in 0's and 1's. Based on these bits, the transformation matrix was updated by applying the Neighborhood Discriminant Hashing (NDH) method [1].

For efficient image retrieval process, genetic optimization algorithm is implemented [6]. Genetic algorithm is considered as an evolutionary algorithm which is determined as a heuristic search technique. This algorithm undergoes five process as initialization, selection, crossover, mutation and termination [26]. These five terminologies undergoes to obtain the optimized feature subset. Selection operator selects the parent individuals by considering the fitness value. Crossover proceeds with generating the two new individual sets and mutation operator proceeds as same as the crossover operator. While obtaining the maximum fittest value, the algorithm gets terminated and the optimized feature subset will be obtained.

The following paper is contributed as, Section II describes the survey on genetic algorithm and similarity search. Section III elaborates the methodology used in this paper and Section IV shows the results obtained for applying the genetic algorithm. Finally, Section V concludes the techniques and results obtained.

II. RELATED WORK

The genetic optimization algorithm is survived related to the image retrieval technique. The descriptions for each specific technique is presented by considering the prior work. The survey is divided into two forms as genetic algorithm and similarity search.

A. Genetic Algorithm

Bir Bhanu et al (2009) described Self-optimizing Image segmentation system based on genetic algorithm [7]. The segmentation process was considered as a difficult approach for any automated process and from the survey on many optimization algorithm, it was considered as a major problem in that technique. The hyperspace which has been correlated in the segmentation problem was searched in the genetic algorithm. In this technique, the recognition rate was low interms of precision and recall.

Xiabi Liu et al (2015) proposed feature selection method and genetic optimization algorithm [6]. The fittest features were selected among the various extraction method interms of bag-of-words, wavelet transform method and histogram. They used classifiers as SVM, KNN and Naive Bayes to label the images in the dataset. Based on those classification, the retrieval process was processed. Among these classifiers SVM classifier performed best in classifying the labels among the classes. But KNN and Naive Bayes classifiers performance were lagged comparing with other techniques.

Jiang Suhua et al (2015) described two Dimension Threshold Image Segmentation [8]. Based on Improved Artificial Fish-Swarm algorithm, the segmentation threshold algorithm was improved. Swarm produced the better segmentation result on comparing with the other optimization techniques. But in this algorithm does not produce the optimal solution for segmentation technique.

Feiping Nie et al (2008) used trace ratio criterion for feature selection [9]. Two popular feature selection algorithm as Laplacian score and fisher score are proposed for selection process. The score of the selection technique for the feature subset was directly optimized. For some criterion, the optimal solution was not found.

Li Zhuo and Jing Zheng (2008) designed a genetic algorithm based wrapper feature selection method for classification of hyper spectral images using support vector machine [10]. The feature subsets and SVM kernel parameters were optimized at the same time by using this technique. GA-SVM method was significant in reducing the computational complexity and the classification accuracy was also improved. But this algorithm regretted more computer memory.

Medical image analysis in artificial neural network [11] was proposed by J. Jiang, P. Trundle et al (2010). The basic concept of neural network was showed for easy understanding. Non-rigid registration of medical images were used to register a patient's data to an anatomical atlas. But in general, the neural network was difficult to interpret and analyse the datas. In some cases they define the process of transforming the input to output in simple manner.

Attakitmongcol K and Srikaew A (2005) proposed a new approach for optimization in watermarking by using genetic algorithm [12]. The spread spectrum image watermarking algorithm was proposed by using discrete Multiwavelet transform. The visual quality of watermarked images and robustness of the watermark were generally improved in their performance.

Khaled Loukhaoukha et al (2010) described Multi-objective Genetic Algorithm for Image watermarking based on singular value decomposition and Lifting Wavelet transform [13]. Multiple Scaling Factors were used to achieve the highest robustness without losing watermark transparency. But determining the optimal values for Multiple Scaling Factors was quite difficult and also unpredictable.

B. Similarity search

Kulis and Jain (2009) proposed a learned metric method to efficiently index into a large database [14]. They used the mahalanobis distance to calculate the nearest neighbor among them for the searching process. This distance calculation predicts the similarity and dissimilarity among the images. Fast approximate similarity search with learned metrics was introduced to construct randomized hash functions that integrate among the partially labeled data and the paired constraints. The main idea of this technique was to learn a Mahalanobis metric based on the training examples, and to encode the learned information into randomized hash functions.

Li and Liu (2015) considered the classification problem in the multi-subspace setting using sparse representation techniques [15]. They presented the extracted information in the dictionary with all the training data. In the dictionary, the trained datas were represented into blocks and where each blocks were named as labels. The goal of this technique was to find a representation of a class which was initialized with the minimum number of blocks from the dictionary. They also formulated the problem of classification by using two non-convex optimization programs to solve the redundancy among the dictionary samples.

Lowe (2004) determined a method for extracting distinctive invariant features from images which has been used to perform reliable matching between different views of an object or scene [16]. This feature provided a robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. These were invariant to image scale and rotation. The features were highly different, in the sense that a single feature can be correctly matched with high probability against a large scale database of features from many images. Using a fast nearest-neighbor algorithm, the recognition proceeds by matching individual features to a database of features from known objects.

Andoni (2009) constructed the Nearest Neighbor (NN) problem which was used to find the similarities among the large scale database [17]. For instance, it formed the basis of a widely used classification method in machine learning to label for a new object, and the most similar labeled object and copy its label. Geometric notions was used to represent the objects and the similarity measures. Using this approach, a new efficient NN algorithm was proposed for a variant of the edit distance, which achieved a double-logarithmic approximation. The similar images was found by considering the distance estimation. This produced the exponential improvement over the lower bound on the approximation achievableTo complement this algorithm, they proved lower bounds on NN data structures for the Euclidean distance which estimates the distance among the parallel images in the database.

III. METHODOLOGY

The genetic algorithm is a search heuristic which shows the process of natural evolution. This technique generates the useful solutions for the optimization problem [6]. Each solution is important for the next generation. It selects the feasible solution among the initialized results. In a genetic algorithm, a population (called chromosomes or the genotype of the genome), which is determined as candidate solutions (called individuals, creatures or phenotypes) for an optimization problem, which is evolved towards the better solutions. The flow for GA is represented in fig 1.

Normally, the binary string is represented as 0's and 1's. For generating the binary string, the mean is calculated for the feature vector of each individual in the population. The greater values are declared as 1 and remaining values are declared as 0.

Termination is achieved by attaining the maximum fittest value or attaining the satisfactory value which is assigned by the user [6], [18]. Until achieving the fittest value, the iteration goes on and finally obtains the optimized feature subset (i.e., dimensionality reduction).

A. Population Initialization

The population initialization is processed to initialize the individuals in the population. Population is considered as set of individuals where each individual is considered as a chromosome. The population is initialized by p decimal values representing each individual.

B. Selection

Selection operator selects the parent individual by considering the mean (i.e., fitness value) for each individual in the class. The top two fittest value are selected as a parent individual. The probability for selecting the parent individual is described as,

$$P(A/B) = P(A) \times P(B) \quad (1)$$

Where, P(A) represents the feature value of first image and P(B) represents the feature value of second image.

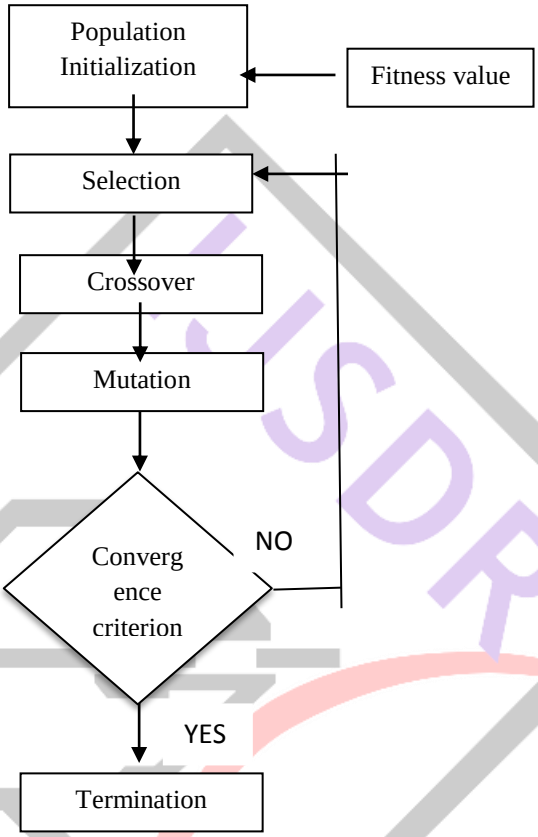

FIGURE 1: FLOW DIAGRAM FOR GENETIC ALGORITHM

C. Crossover

Crossover operator swaps the binary bit to generate two new offspring with the presence of parent individuals. Crossover undergoes three types as one point crossover, two point crossover and uniform crossover.

In one point crossover, the single point in one individual is replaced with the other individual which is represented in fig 2. From the parent individual one point is selected and where the selected point's remains the same and the rest of the points are swapped with other individual.

In two point crossover, two points are decided to change from one individual. Two points will be swapped among the individuals which is represented in fig 3. Two points are selected in the parent individuals and the selected points remains the same where others points are swapped among them.

In uniform crossover, the points will be randomly selected and it will be swapped among the individuals. In the parent individuals the points are randomly chosen where there is no restriction for selecting the points.

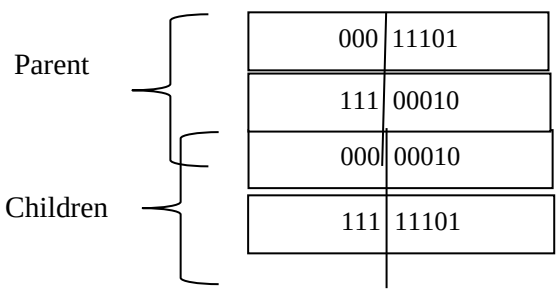

FIGURE 2: ONE POINT CROSSOVER

D. Mutation

This operator proceeds as same as the crossover operator. In this operator, it changes one bit from the generated offspring to generate the population for the next generation.

E. Termination

The algorithm gets terminated while attaining the maximum fittest value among the adjacent generation or attaining the satisfactory fitness value.

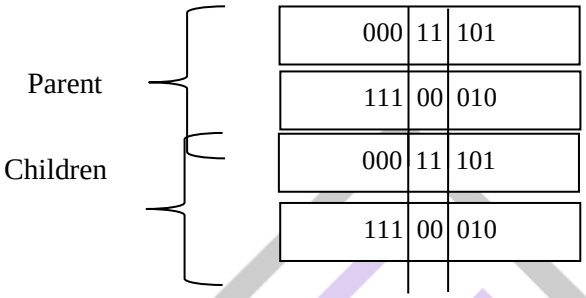

FIGURE 3: TWO POINT CROSSOVER

IV. EXPERIMENTS AND RESULTS

A. Datasets

Four publically available datasets are considered i.e., AR, MIR Flickr, NUS-Wide and ORL datasets. AR dataset [19], contains about 4000 colored images of two sessions as female and male. In each session, it contains about 13 images of dimension 120 × 165. The images are shown in RAW and BMI file format.

MIR Flickr dataset [20], is assigned with 25,000 images with 1386 tags and 38 labels from the flickr website. It provides the clarity images with high quality and resolution.

NUS-Wide dataset [21], contains 2,70,000 images with 1000 frequent tags which has been collected from the flickr benchmark.

ORL dataset [19], contains about 400 images with different samples. The images in this dataset are correlated with different lighting conditions and facial expressions. Each session is assigned with 10 images with the dimension 92 × 112. The images are represented in PGM format.

B. Feature Extraction

The feature extraction is used to extracts the values from the images. Four extraction methods are commonly used as GIST, Pixel, BoW and SIFT. These features converts the raw data to the decimal values.

GIST [22], which is assigned to the AR dataset. This feature extracts the values by considering the intensity of the image which is associated with 0 and 1.

Pixel feature [23], assigned to the MIR Flickr dataset. It estimates the values by calculating the mean for the pixel value of the image [25].

BoW feature [21], assigned to the NUS-Wide dataset. This feature estimates the value interms of cluster formation. For each cluster, new codeword will be generated.

SIFT feature [16], assigned to the ORL dataset. It estimates the values by detecting the keypoints in the image and by performing localization and orientation for the detected keypoints.

C. Results

The query image is converted to the gray image for easy convergence where, the ray image is represented in single band. For the converted gray image, the four extraction methods are applied to convert the high dimensional data into low dimensional form. The matrix is created for the whole dataset with its assigned dimension.

TABLE 1: PRECISION FOR FOUR DATASETS

Precision		
Dataset/Method	**NDH**	**GA**
AR	0.75	0.88
MIR Flickr	0.67	0.73
NUS-Wide	0.65	0.77
ORL	0.82	0.89

TABLE 2: RECALL FOR FOUR DATASETS

Recall		
Dataset/Method	**NDH**	**GA**
AR	0.8	0.88
MIR Flickr	0.71	0.75
NUS-Wide	0.70	0.79
ORL	0.84	0.9

To normalize into the optimized feature subset, the genetic optimization algorithm is incorporated. This algorithm obtains the feature subset by eliminating the redundant features which undergoes five terminologies.

Genetic algorithm is generally assigned with the stopping criterion. It is normally assigned as an iterative process. While attaining the maximum fittest value, the algorithm gets terminated. The results are shown in table 1 and 2.

D. Performance Evaluation

Two performance metrics are considered i.e., precision and recall. Precision represents the irrelevant images which has been obtained for the query image. Recall indicated the relevant images which is correctly related to the query image.

V. CONCLUSION

Due to rapid development of web usage technology, the usage of users and information shared by the user are also high. So, the retrieval process became more complicated on comparing with earlier days. Many optimization techniques were used but, the recognition rate was low in prior NDH method. In this method, the recognition rate is 90% improved interms of precision and recall.

REFERENCES

[1] Jinhui Tang and Zechao Li, "Neighborhood Discriminant Hashing for Large-Scale Image Retrieval," Image Processing, vol. 24, no. 9, sep 2015.

[2] Data mining- "Similarity Search", url: https://similarity search.com

[3] Image retrieval- "Feature Extraction", url: https://feature extraction.com

[4] Image mining- "Feature selection", url: https://feature selection.com

[5] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing for scalable image search," in Proc. IEEE 12th Int. Conf. Comput. Vis., Sep./Oct. 2009, pp. 2130–2137.

[6] Xiabi Liu, Ling Ma and Li Song, "Recognizing Common CT Imaging Signs of Lung Diseases Through a New Feature Selection Method Based on Fisher Criterion and Genetic Optimization," IEEE Journal of Biomedical And Health Informatics, Vol. 19, No. 2, March 2015.

[7] Bir Bhanu, Sungkee Lee and John Ming, "Self-optimizing Image segmentation system using a genetic algorithm," 2009.

[8] Jiang Suhua¸ Liu Chunqiang and Wang Dongdong, "Two Dimension Threshold Image Segmentation Based on Improved Artificial Fish Swarm Algorithm," International Conference on Chemical, Material and Food Engineering (CMFE-2015).

[9] Feiping Nie, Shiming Xiang, Yangqing Jia, Changshui Zhang and Shuicheng Yan, "Trace Ratio Criterion for Feature Selection," Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008).

[10] Li Zhuo, Jing Zheng, "A Genetic Algorithm based Wrapper Feature Selection Method for Classification of Hyperspectral Images Using Support Vector Machine," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7. Beijing 2008.

[11] J. Jiang, P. Trundle, J. Ren, "Medical image analysis with artificial neural networks," Computerized Medical Imaging and Graphics 34 (2010) 617–631.

[12] Attakitmongcol K and Srikaew A, "A new approach for optimization in watermaking by using genetic algorithm," IEEE Transaction on Signal Processsing," DOI: 10.1109, Sep 2014.

[13] Khaled Loukhaoukha, Jean-yves Chouinard and Mohamed Haj Taieb, "Multi-objective Genetic Algorithm for Image watermarking based on singular value decomposition and Lifting Wavelet transform," Image and Signal Processsing, July 2010.

[14] Kulis B. and Jain P. (2009), 'Fast similarity search for learned metrics', IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 12, pp. 2143–2157.

[15] Li Z. and Liu J. (2015), 'Robust structured subspace learning for data representation', IEEE Trans. Pattern Anal. Mach. Intell., doi: 10.1109/TPAMI.2015.2400461.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.

[17] Andoni A. (2009), 'Nearest neighbor search: The old, the new, and the impossible', Massachusetts Inst. Technol., Cambridge, MA, USA.

[18] T. Sowmi and E. SaravanaKumar, "A survey on Fisher Criterion based Genetic Algorithm for Feature Selection Method," South Asian Journal of Engineering and Teachnology, Vol.2,no.12 (2016) 1-8.

[19] Marryam Murtaza, Muhammad Sharif, Mudassar Raza, and Jamal Hussain Shah (2014), 'Face Recognition Using Adaptive Margin Fisher's Criterion and Linear Discriminant Analysis (AMFC-LDA)', The International Arab Journal of Information Technology, Vol. 11, No. 2.

[20] Huiskes M.J. and Lew M S. (2008), 'The MIR Flickr retrieval evaluation', in Proc. ACM Int. Conf. Multimedia Inf. Retr., pp. 39–43.

[21] Chua T.S. and Tang J. (2009), 'NUS-WIDE: A real-world Web image database from National University of Singapore', in Proc. ACM Int. Conf. Image Video Retr., Art. ID 48.

[22] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in Proc. ACM Int. Conf. Multimedia Inf. Retr., 2008, pp. 39–43.
[23] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 11, pp. 1958–1970, Nov. 2008.
[24] T. Sowmi and E. Saravana Kumar, "A Survey on Similarity Search for Large Scale Database," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2015.
[25] Karen Glocer, Damian Eads and James Theiler, "Online Feature Selection for Pixel Classification," Appearing in Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.
[26] Mohd Saberi Mohamad, Safaai Deris, Safie Mat Yatim and Muhammad Razib Othman, "Feature Selection Method using Genetic Algorithm for the Classifcation of Small and High Dimension Data," First International Symposium on Information and Communications Technologies. October 7-8, 2004. Putrajaya, Malaysia.