ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

A Survey on Similarity Search for Large Scale Database

T. Sowmi[1], E. Saravana Kumar[2]

P.G Scholar, Dept of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India[1]

Associate Professor, Dept of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India[2]

Abstract: Data mining is the process of extracting the needed information and knowledge from the database based on the user's expectance. In Image mining, Similarity search is one of the ongoing research fields for efficient image retrieval. For similarity search, many algorithms based on hashing and genetic were implemented. The aim of these algorithms is to retrieve the relevant images which are similar to the given query image based on feature extraction method. It estimates the fitness value for binary bit generation and will obtain the optimal solution for image retrieval. Among the reviewed papers four datasets were commonly used. The datasets used are MIRFLICKR, CIFAR-10, NUS-WIDE and SIFT-1M which provides clarity for image and specificity for querying. Mostly the implementation was conducted in Matlab which uses Fortan for version 1 and C for commercial use.

Keywords: Similarity search, Hashing, Genetic algorithm, Feature extraction method.

I. INTRODUCTION

Data mining is the analytic process designed to extract the needed information, pattern and/or knowledge from the large scale database. It has presented a major role in society as well as in the information industry due to the availability of large amount of data. Most of the industries are based on data mining techniques and their ultimate goal is to predict the relevant information.

Image mining is considered as a type of data mining. It extracts the relevant pattern and/or information which are not explicitly found in the images. Nowadays, most of the social media networks such as Facebook, Google+, Twitter etc., contains large amount of documents and images which has been uploaded and shared by users. Soretrieving the relevant information is very difficult task particularly, retrieving the relevant images for a given query image is the challenging task and is becoming an ongoing research issue. Therefore, similarity search in large scale datasets has gained one of the main research issues among multimedia communities. Similarity search is defined as identifying the similar samples in the database for the given query sample. Many similarity search algorithm has been defined refer [2].To improve the efficiency of retrieval process, hashing based techniques is used which converts the high-dimensional features into low-dimensional compact binary codes. Recently, many types of hashing function were developed for the consideration of Similarity search. For preserving the local discriminant information and for efficient learning of hashing functions a novel Neighbourhood Discriminate Hashing function was developed based on feature extraction method[1].Most commonly used feature extraction methods are SIFT [4], GIST[5], pixel[3] and BoW [4]. This algorithm predicted the optimal solution forretrieval process by estimating the Objective function and by updating the transformation matrix in terms of probability estimation.

To improve the performance and recognition rate, Fisher Criterion based Genetic Algorithm will be implemented [14], [15]. The genetic algorithm estimates the fitness value to generate binary strings as 0s and 1s based on selection, crossover and mutation operator. Then the Fisher criterion algorithmestimates the mean and covariance value and then estimates the Eigen decomposition value for obtaining the optimal solution.

The purpose of this literature survey is to review the image retrieval process based on Genetic and hashing algorithm for the efficiency of retrieval process. In this way this survey will be carried out to improve the efficiency, Computational complexity and performance metrics in terms of precision (i.e., irrelevant images) and Recall (i.e., relevant images) by varying the bits.

The remainder of this paper is arranged as follows. Related work is described in Section 2. We elaborate the algorithm in Section 3 and 4. Datasets based review is described in Section 5. How the research is analysed is described in Section 6. Observation for the research questions are carried out in Section 7. Finally, the discussion and conclusion is analysed in Section 8 and 9.

II. RELATED WORK

In this literature, the image based retrieval techniques has been studied widely in occasion of Genetic algorithm, Similarity search and Hashing techniques are as follows and refer table 1.

B. Kulis and P. Jain (2009) have proposed a fast similarity search [6] based on learned metrics to estimate the index values (i.e., pixel values) and Mahalanobis distance for trained images. This technique improved the accuracy of search.

Z. Li and J. Liu (2015) have considered the Robust structured subspace learning [7] for data representation in

DOI 10.17148/IJARCCE.2015.4923

terms of classification using sparse representation (i.e., nonzero portion). They also exploited the dictionary for each class in terms of block structure.

A. Andoni (2009) has proposed the Nearest Neighbor search algorithm [8] to estimate the Euclidean distance for both query images and images in the database. This algorithm solved the exact nearest neighbor search problem for image retrieval. But it was not suitable for modern datasets and it assumes more time for largest datasets.

Jianfeng Wang and Jingdong Wang (2013) have proposed Order preserving hashing method [9] for approximate nearest neighbor search to preserve the hash functions for similarity search and the main aim of this algorithm was to align the similarity orders (i.e., Ascending order) both in hamming space and original space. This algorithm was not limited and generalized to other hash functions.

W. Liu and J. Wang (2012) have developed a novel kernel based supervised hashing method [10] to predict the similar and dissimilar data pairs which does not include any semantic labels and possibly achieved limited training cost for high quality hashing. This algorithm also estimated the hamming distance (i.e., compact binary codes) for all images in the database.

David G. Lowe (2004) has presented a method for extracting distinctive invariant features [4] for images to perform reliable matching in different views of object. This method has not considered any consistent measures. So it discards image information and limited the descriptors to be used. The features are changed against in 3D viewpoint for non-planar surfaces.

Chenxia Wu and Jianke Zhu (2013) have studied the effective semi-supervisedNonlinear hashing method [11] to capture the underlying relationship among data points. But some linear hashing method will not effectively reflect the data points. The dimensionality of the computation was independent on both original space and hamming space.

Bin Xu andJiajun Bu (2013) have introduced harmonious hashing algorithm [12], which aimed to minimize the information loss due to feature extraction. This model is not sensitive to the parameters. The minimum distance value was kept on the top dimension and the maximum values were not used.Jun wang and Sanjivkumar (2012) have presented semi-supervised hashing [13] to minimize the empirical error (i.e., loss of information) over both labelled sets and unlabeled sets.

Dong Xu and Shuicheng Yan (2007) have proposed fisher analysis and gait recognition [14] which aimed to select the similar and discriminant information based on feature extraction method. This algorithm directly handles the 2-D images into the gray scale images for easy convergence.

Xiabi Liu and Ling ma (2015) have proposed Fisher criterion and Genetic optimization method [15] which was shortly termed as FIG. This algorithm found the optimal feature subset from the candidate data features. The selected features were incorporated with five mining techniques to classify the regions.

Chih-chin Lai and Ying chuanchen (2011) have considered Interactive genetic algorithm (IGA) [16] to reduce the gap between the retrieval results and user's expectation. This method considered mean value, standard deviation and image bitmap of color image as a feature extraction method. IGA employed to satisfy the user's need by identifying the images based on their requirements.

Most of the researches researched many hashing functions for similarity search and Genetic algorithm based approaches to retrieve the relevant images for a query image from a collection of images. In order to improve the performance and efficiency, the above contributions are extended further by suggesting some more features in NDH algorithm.

III. GENETIC ALGORITHM

A genetic algorithm is a search based Meta heuristic approach [16] which undergoes natural selection process (i.e., fittest selection). This algorithm belongs to the class of Evolutionary algorithm (EA) which generates exact solution for optimization and search problems using natural evolution techniques such as mutation, selection and crossover terms. The optimization problem is evolved towards the better solutions due to the availability of candidate solutions. Each candidate solutions can be altered and mutated traditionally and it is represented in terms of binary string as 0s and 1s.

The evolution starts for each individual in the candidate solutions. It is considered as an iterative process whereeach iteration is termed as generation. The fitness value for every individual is evaluated foreveryiteration and this represents the value of the objective function in the optimization based problem. The estimated value will be in binary form.

The genetic algorithm is based on genetic representation of the solution domain and fitness value. Once the genetic representation and fitness value is derived, the genetic algorithm undergoes mutation, crossover, inversion and selection operators to generate binary bit for the data matrix. The main key idea of selection operator is to estimate the better individual based on fitness. The fitness value is estimated based on objective function or subjective judgment.Two individuals are chosen from the candidate solution using the selection operator. The bit strings is randomly chosen along with the crossover site and the values of the two strings are exchanged. By recombining the portions of good individuals, this process creates even better individuals. Mutation operator inhibits the premature convergence.

IV. FISHER CRITERION ALGORITHM

Fisher criterion algorithm is considered as a classification method [14] which projects the high dimensional data into one dimensional space. This projection maximizes the mean value within the class by minimizing the variance of the class. The projection for fisher criterion algorithm is defined by,

$$(W) = \frac{|m_1 - m_2|^2}{s_1^2 + s_2^2}$$

DOI 10.17148/IJARCCE.2015.4923

Year	Title	Methodology	Performance metrics	Datasets	Experimental method
2008	Spectral Hashing	Similar samples contains same binary codes which has been detected by semantic hashing	Precision and Recall	80 million images from the Internet	Study
2009	Kernelized Locality-Sensitive Hashing for Scalable Image Search	Optimal similarity search of retrieving items with (1 + £) times	Recall rate	Caltech 101 and Photo Tourism	Matlab implementation
2012	Spherical Hashing	Locality-Sensitive Hashing (LSH)	Precision and Recall	GIST-1M-384D, GIST-1M-960D and GIST-75M-384D	i7 X990 with 24GB main memory and X5690 with 144GB main memory
2012	Supervised hashing with kernels	Kernel-Based Supervised Hashing (KSH)	Precision, Recall and lookup success rate	CIFAR-10 and Tiny-1M	2.53 GHz Intel Xeon CPU and 48GB RAM
2012	Semi- supervised hashing for large scale search	Approximate nearest neighbors (ANN).	Precision and Recall	CIFAR10, Flickr image collection-NUS-WIDE and the 80 million tiny images	Lenovo workstation with 3.16 GHz Quad Core CPU- Matlab implementation
2012	Super-Bit Locality-Sensitive Hashing	Sign-random-projection locality-sensitive hashing (SRP-LSH)	Super-Bit Depth and Code Length	Photo Tourism and MIR-Flickr	Study
2013	Order preserving hashing for approximate nearest neighbor search	Novel ANN search approach order preserving hashing (OPH)	Recall, code length and Precision	GIST1M	Intel Xeon CPU of 3.33GHz and 24GB memory-Matlab implementation
2013	Semi-Supervised Nonlinear Hashing Using Bootstrap Sequential Projection Learning	Effective semi-supervised hashing method under the framework of regularized learning-based Hashing	Mean Average Precision (MAP), precision within Hamming radius and recall	MNIST, ISOLET, USPS, Caltech101, SIFT1M and PATCH1M	Matlab implementation

Table 1: List of Hashing Function Based Paper

Where, m is the mean of two classes, $s_1$and s_2 is the variance of two classes. In signal theory, this algorithm is considered as Signal to Interference ratio. This algorithm forms the optimal solution by maximizing this criterion.The projected values are spread out as much as possible while comparing with variance. The linear combination $Z = a^T X$ will be estimated where between-class variance B is maximized with respect to the within-class variance W.

By linear combination, between-class variance is estimated as $a^T Ba$and within-class variance is estimated as $a^T Wa$. Then Fisher optimization becomes,

$$\max_a \frac{a^T Ba}{a^T Wa}$$

To optimize the above criterion Eigen decomposition is estimated. It can be determined as,

$$B^* = (W^{-\frac{1}{2}})^T BW^{-\frac{1}{2}} = D_W^{-\frac{1}{2}} V_W^T B V_W D_W^{-\frac{1}{2}}$$

V. DATASETS

Among the considerations of reviewed papers, most of the implementation where conducted in Matlab with publically available datasets for the efficiency. For image retrieval process, many datasets are publically available. Most commonly used datasets are MIRFLICKR, CIFAR-10 and NUS-WIDE which provides clarity for image retrieval.

Mark J. Huiskes and Michael S. Lew (2008) have presented that the MIRFLICKR datasets [17] contains about 25000 images which are used for research purposes and it also represents a real community of users. But all the images in this website was not copyright-free and some images was not clear in clarity.

Antonio Torralba and Rob Fergus (2008) have studied 80million tiny images[3] with the combination of simple technique. The clarity of the image was based on search engine used and the image specificity was described for querying. They also described that deriving the index values for large datasets was difficult task but, this intimation was not described in this paper.This dataset contains social media logo, natural scenarios etc.,

T.S. Chua andJ.Tang have studied a real world web image [18]database from National University of Singapore. This dataset was created for media search which provided clarity and indication for querying. The images recovered from this dataset were based on medical field images which provided clarity among the images.

The above described contribution expels the clarity and specificity of images for retrieval. So these three datasets were commonly used for the implementation.

VI. RESEARCH METHOD

This study is considered under a Systematic Literature Review (SLR) for similarity search from large scale database. A research questions plays a major role in the survey and it produces clarity for this survey. The questions related to similarity search and image retrieval are described as follows.

A. RESEARCH QUESTIONS

1) What are the available hashing techniques for image retrieval?
   Motivation: Hashing techniques were used to obtain binary bit and transformation matrix for easy convergence.
2) Why it is considered as a research issue among multimedia communities?
   Motivation: Due to the enormous collection of images, retrieving the relevant images is becoming a critical task.
3) Why Matlab language is suitable for image based retrieval process?
   Motivation: This language is particularly concerned for numerical computation and data visualization. It is easier for the conversion of image to matrix format.
4) What are the performance metrics considered for similarity search?
   Motivation: Among the reviewed papers most of the implementation was based on varying precision and recall metrics.
5) What is the need for using hashing function?
   Motivation: For the conversion of decimal values to binary bit, the hashing function is used which makes the retrieving process easier and improves the performance.

VII. OBSERVATION

From the reviewed papers, the answers for the research questions are observed. It is presented below interms of retrieval process and Matlab.

1) What are the available hashing techniques for image retrieval?

For Image retrieval, many hashing techniques are available. Some frequently followed techniques are Semantic Hashing [19], Isotropic Hashing [20], Spectral Hashing [21] etc., These hashing techniques generates binary bits and determines the transformation matrix. But none of these techniques derived optimal solution for optimization problem. To generate the optimal solution, a novel NDH method was used which predicts the objective and updates transformation matrix.

To improve the recognition rate and performance, genetic algorithm and fisher criterion are considered which maximizes the mean value by minimizing the variance and estimates the eigen value for optimization purpose.

2) Why it is considered as a research issue among multimedia communities?

Due to the rapid development of digital technologies, many photos sharing website such as FLICKR, Google+, Facebook etc., contains many images which has been shared and uploaded by the users. Comparing to earlier days, at present 6million users are using photo sharing website and/or social media networks. So, the availability of images shared by the users has also been increased.

Inspite of that, to retrieve the relevant images for the given query image has become the major research issue among multimedia communities. This retrieving technique is not only particularly used for photo sharing website. It is commonly applicable for alllarge databases which were concerned with images.

3) Why Matlab language is suitable for image based retrieval process?

Matlab is determined as Matrix Laboratory [22] which is designed by Cleve Moler. It is considered as a high level language and a multi-paradigm numerical computing environment. It is concerned particularly for the conversion of image to matrix format. As well as this language consists of many inbuilt functions which are applicable for the image retrieval process and plotting the graph for analyzing the performance metrics is also much easier. The variables in Matlab were considered as array which was determined as Structure array. So while comparing with other language, Matlab is best suited for Similarity search. The coding is written in C for commercial and in Fortan for version 1 with extension m.

4) What are the performance metrics considered for similarity search?

From the reviewed paper, mostly the considered parameters are Precision (i.e., irrelevant images) and Recall (i.e., relevant images). Precision determines the irrelevant images from the retrieved samples and Recall determines the relevant images which are equivalent to the query image. The comparison with the state of the art

DOI 10.17148/IJARCCE.2015.4923

techniques were also estimated interms of the performance metrics and the graphs was plotted by increasing the number of hashing bits to determine the precision curve. By increasing the hashing bits, the information of the image were increased. When the image information is increased the irrelevant images will be detected and retrieved.

5) What is the need for using hashing function?

When the matrix was created for the images in the database, it produces a decimal value based on feature extraction method which was very difficult to predict the relevant images. So to convert the decimal values to binary bit, the hashing function was used. Particularly LSH [23] function is used which reduces the dimensionality of images. The binary bits were generated based on hashing bits and hashing function. Linear hashing function was used normally for the generation of hashing bit which was more efficient for image retrieval. This function was commonly used in many hashing based methods.

VIII. DISCUSSION

Image retrieval is considered as one of the challenging task among the research communities. Because retrieving the relevant images for the query image based on feature extraction method is not as much easier.

Nowadays, normally collection of images is increasing day by day due to the development of digital technologies. So searching and retrieving the relevant imagesfrom the large setsis becoming a critical task. To overcome this problem, hashing based image retrieval techniques were followed. Many hashing based methods were described in earlier days for similarity search and efficient image retrieval.

To obtain the optimal solution for image retrieval Jinhui Tang(2015) proposed a novel Neighborhood discriminant Hashing method which solved the optimization problem by updating the transformation matrix. This algorithm created the data matrix based on feature extraction methods. The methods mostly followed were GIST, SIFT, pixel and BoW. Gist feature reduces the dimensionality and estimated the values based on orientation (i.e., block wise) and frequency which does not require any segmentation. SIFT feature described that theimage key points (i.e., angle orientation) are extracted from a set of reference images and the points were stored in a database. Pixel feature estimated the mean values for images in the database.Bag-of-words feature used dictionary of words in the hashing trick, where words are directly mapped to indices based on hashing function.The matrix creations were estimated with decimal values.

For easy convergence, Locality-Sensitive Hashing approach was derived for the conversion to binary bit. This approach reduces high-dimensional data to low dimensionality. It hashes image as an input so the similar images were mapped to the same buckets with high probability. This method differs from other hash functions such as conventional and cryptographic hash functions. Because it was mainly used to maximize the probability and to solve the Nearest Neighbor (NN) searches.

Probability was estimated for an assumption using the treated neighbors both in hamming space and original space. Finally NDH method was applied to derive the objective function for the conversion of discrete to continuous values and also updated the transformation matrixuntil it reaches certain convergence criterion for efficient image retrieval. If the convergence was satisfied then the relevant images were retrieved else the process was again continued from probability estimation.

To improve the recognition rate and performance analysis, the Fisher criterion based genetic algorithm is referred. This algorithm derives the optimal solution for image retrieval by estimating the fitness value for binary bit generation by using genetic algorithm and generates the mean and variance for an optimization problem by fisher optimization criterion algorithm.

IX. CONCLUSION

Due to the enormous development of digital technologies, storage of images has been increasing day by day. So to retrieve the relevant images many hashing and genetic based algorithm were implemented previously. Among the reviewed papers, many implementations were conducted in Matlab due to the availability of many inbuilt function for conversion. In order to improve the recognition rate and performance, specifically Fisher criterion based genetic algorithm is consideredby evaluating four datasets in Matlab and the performance metrics will be estimated.

REFERENCES

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

[2] A. Andoni, “Nearest neighbor search: The old, the new, and the impossible,” Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.

[3] A. Torralba, R. Fergus, and W. T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” IEEETrans. Pattern Anal. Mach. Intell., vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

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

[5]A. Torralba, R. Fergus, and Y. Weiss. “Small codes and large image databases for recognition,” In CVPR, 2008.

[6]B. Kulis, P. Jain, and K. Grauman, “Fast similarity search for learned metrics,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 12, pp. 2143–2157, Dec. 2009.

[7] Z. Li, J. Liu, J. Tang, and H. Lu, “Robust structured subspace learning for data representation,” IEEE Trans. Pattern Anal. Mach. Intell., Feb. 2015, doi: 10.1109/TPAMI.2015.2400461. [Online]. Available:http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=7031960.

[8] A. Andoni, “Nearest neighbor search: The old, the new, and the impossible,” Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.

[9] J. Wang, J. Wang, N. Yu, and S. Li, “Order preserving hashing for approximate nearest neighbor search,” in Proc. ACM Int. Conf. Multimedia, 2013, pp. 133–142.

[10]W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, “Supervised hashing with kernels,” in Proc. IEEE Int. Conf. Comput. Vis. PatternRecognit., Jun. 2012, pp. 2074–2081.

DOI 10.17148/IJARCCE.2015.4923

[11] C. Wu, J. Zhu, D. Cai, C. Chen, and J. Bu, "Semi-supervised nonlinear hashing using bootstrap sequential projection learning," IEEE Trans.Knowl. Data Eng., vol. 25, no. 6, pp. 1380–1393, Jun. 2013.

[12] B. Xu, J. Bua, Y. Lin, C. Chen, X. He, and D. Cai, "Harmonious hashing," in Proc. 23rd Int. Joint Conf. Artif. Intell., 2013, pp. 1820–1826.

[13] J.Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for largescale search," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 12, pp. 2393–2406, Dec. 2012.

[14] Dong Xu and Shuicheng Yan. "Marginal Fisher Analysis and its Variant for Human Gait Recognition and Content- Based Image Retrieval," Image processing, Vol. 16, No. 11, 2007.

[15] Xiabi Liu and Ling ma, " Recognizing Common CTImaging Signs of Lung Diseases Through a New feature Selection Method Based on Fisher Criterion and Genetic Optimizatiom," Health Informatics, Vol. 19, No. 2, 2015.

[16] Chih-chin Lai and Ying chuanchen, "A User- Oriented Image Retrieval System Based on Interactive Genetic Algorithm," Instrumentation and Measurement, Vol. 60, No. 10, 2011.

[17] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," inProc. ACM Int. Conf. Multimedia Inf. Retr., 2008, pp. 39–43.

[18] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE:A real-world Web image database from National University of Singapore," in Proc. ACM Int. Conf. Image Video Retr., 2009, Art. ID 48.

[19] R. Salakhutdinov and G. Hinton, "Semantic hashing," Int. J. Approx. Reasoning, vol. 50, no. 7, pp. 969–978, 2009.

[20] W. Kong and W.-J. Li, "Isotropic hashing," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1646–1654.

[21] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in Proc. Adv. Neural Inf. Process. Syst., 2008, pp. 1753–1760.

[22] Cleve Moler, "Matlab Programming", in wiki books, 2015.

[23] P. Indyk and R. Motwani, "Approximate nearest neighbours: Towards removing the curse of dimensionality," in Proc. ACM Symp. Theory Comput. 1998, pp. 604–613.

DOI 10.17148/IJARCCE.2015.4923