Medical Image Retrieval Using Manifold Ranking with Relevance Feedback

Pooja Soundalgekar[1] Mukta Kulkarni[2] Divija Nagaraju[3] Dr. Sowmya Kamath[4]

Department of Information Technology
National Institute of Technology Karnataka, Surathkal
Mangalore, India
{[1]pooja27ms,[2]mukta3396,[3]divijanagraju}@gmail.com
[4]sowmyakamath@nitk.edu.in

***Abstract*—Medical image retrieval (MedIR) is a challenging field in Visual information retrieval, due to the multi-dimensional and multi-modal context of the underlying content. Traditional models do not take the intrinsic characteristics of data into consideration and have achieved limited accuracy in application to medical images. Manifold Ranking (MR) is a technique that can be used in further optimizing precision and recall in MedIR applications as it ranks items by traversing a dynamically constructed content-specific information graph. In this paper, a MedIR approach based on Manifold Ranking is proposed. Medical images being multi-dimensional, exhibit underlying cluster and manifold information which enhances semantic relevance and allows for label uniformity. Hence, when adapted for MedIR, MR can help in achieving large-scale ranking across datasets as is the case in most medical imaging applications. In addition, a relevance feedback mechanism was also incorporated to support a learning based system. We show that MR achieved significant improvement in retrieval results with relevance feedback as compared to the Euclidean Distance (ED) rankings. This showcases the importance of analyzing the inherent latent structure in medical image data for better performance over traditional methods.**

***Index Terms*—Manifold Ranking, Image Retrieval, Relevance Feedback, Medical informatics**

I. Introduction

Due to the availability of advanced, state-of-the-art software and hardware, there has been rapid advancement in the field of imaging technology, especially in the medical domain. The resulting large volumes of image data have made the process of searching efficiently and accurately for a targeted set of images increasingly difficult, thus making it a problem of significant research interest. Currently, content-based image retrieval (CBIR) and keyword based querying are the most popular methods in MedIR. In text/keyword based querying, traditional database systems and text based annotations stored along with each image are leveraged for retrieval. The approach involved in CBIR utilizes image-level features like color, shape and texture for identifying relevant images for retrieval, in reference to a given query image. Even though text based retrievals are fast, they are only reliable when images are sufficiently annotated. In many situations, results generated can be irrelevant due to ambiguous, low quality annotations or lack thereof. The Picture Archival and Communication (PACS) System [2], [3] was a significant effort in this direction, with a focus on effectively storing, retrieving and transmitting images. However, a major limitation is that these kind of techniques rely only on keywords and associated text annotations stored along with the image. This has led to the development of and subsequent popularity of CBIR systems specifically dealing with MedIR.

CBIR systems focus on capturing those features of an image dataset without relying on any external information like metadata associated with images. Techniques for retrieving images from online databases based on color, shape and texture were studied in the Query Image by Content (QBIC) project [1], which reported several approaches that achieved good results. Consequently, most CBIR systems make use of color and texture as features for producing relevancy rankings. However, a significant hurdle is that fact that most medical images are grayscale, hence, color cannot be used as a prominent feature. But, in medical images, image texture and shape are crucial which need to be effectively captured.

Content based image retrieval was popularized by the ASSERT system [10] that was developed in 1999 for automating indexing of high-resolution computed tomography (CT) scans of lungs in medical imaging. By utilizing the grayscale and texture attributes from the image co-occurrence matrix for characterization of the dataset, the system allowed extraction of pathology-bearing regions in lung scans. The web-based IRMA (Information Retrieval for Medical Applications) [11] system uses a submitted X-ray image for querying, for discovery of similar images from the database. Here, the image retrieval process is broken down to seven steps where every step represents a hierarchy of image abstraction, that signifies a high level understanding of the image content. To compute the closest image for a given query, the local textures and similarity measures are incorporated. Another technique, Flexible Image Retrieval Engine (FIRE) [12] includes non-medical datasets like photographic databases along with different kinds of medical datasets. It utilizes feature engineering concepts to adapt the weights assigned to every feature during retrieval for enhancing performance. Numerous CBIR ranking policies

0-7695-6360-0/18/$31.00 ©2018 IEEE
DOI 10.1109/ICSC.2018.00075

based on IR models like BIM (Binary Independence Model) [6], BM25 (Best Match Okapi 25) [7], Vector Space Model [8] etc., have been developed over the years.

Learning to rank is a retrieval technique that introduced the concept of ranking function optimization for avoiding tuning of large set of relevant features just based on observation. Manifold Ranking [4], [5] is a ranking algorithm that is applied for retrieval tasks based on information graphs generated from low-level image features. MR [9] can achieve better accuracy and performance due to graph construction based on associated image-level information. We therefore consider the application of MR for image retrieval from large-scale medical datasets, and evaluate its effectiveness on standard medical image datasets.

The content of this paper is organized as follows: Section II presents a detailed discussion on the proposed MedIR methodology that adapts the technique of Manifold Ranking for large-scale medical image datasets. The adopted IR model is extended by incorporating a relevance feedback loop, which is discussed in Section III. The experimental evaluation of the proposed approach when applied to standard medical image datasets is presented in Section IV followed by conclusion and future work.

II. Proposed Approach

The proposed technique is an approach based on Manifold Ranking, intended for developing an effective MedIR framework along with relevance feedback for further optimizing the retrieval performance. The processes designed as part of the proposed IR model are described next. The first process deals with preprocessing the given medical dataset for extraction of low-level feature attributes, which are then utilized as graphical data point coordinates for the MR algorithm. The extracted key features help in clearly representing the intrinsic characteristics of the medical images. Since most medical images are gray-scale, it is beneficial to use gray level features and the local texture features to represent the image. The feature vector for the images consists of a histogram of features like mean, variance, skewness, kurtosis and measure of energy (listed below). The features obtained after processing the medical images in the given dataset are considered as dimensions of an image data point and MR algorithm is applied to these images represented by their feature vectors as data points in the graph.

- *Mean*: given by the average values of the all the pixel intensities in an image and denotes the intensity contribution of each pixel in the entire medical image.
- *Variance*: represents the compactness of the image pixel intensities, and is normally used to classify pixels into different regions by finding out how each pixel spreads from the neighboring pixel (or center pixel).
- *Skewness*: Darker and glossier surfaces are positively skewed as compared to light surfaces. Consequently, we can use skewness to understand about image surfaces.
- *Kurtosis*: denotes an interpretation of the image as a combination of noise and resolution.
- *Energy*: describes how gray levels are spatially distributed in the given image.

A. Manifold Ranking (MR)

The next process deals with the construction of the graph structure representing the intrinsic content of the medical images as per the MR algorithm. The basic notation for graphical representation G, with E edges with weight W generated is $G = (V, E, W)$ where $E \subseteq V \times V$. This indicates the interrelation between the vertices and the weights are used to denote the quantitative significance of the pair-wise relation between the vertices. The main objective of this MR based graphical approach is to incorporate a method that captures the significance of each data point with respect to other data points by considering the global and local information reflected by the graph.

Consider a set with data $X = \{x_1, x_2, ...x_n\} \subset R^m$. This is used to build a graph (KNN: K Nearest Neighbor Graph) which is essential to obtain required local information present in a image. The inputs to the algorithm are data points in an adjacency matrix representation format and a query image. The edge weight between two data points $\{i, j\}$ is denoted by w_{ij}. The adjacency matrix $W \in R^{n*n}$ incorporates all the edge weights. Let $w_{ij} = exp[-(d(x_i, x_j))^2/(2 * \sigma^2))]$ be a heat kernel for edge linking x_i and x_j, else $w_{ij} = 0$ [9].

Let $d(x_i, x_j)$ defined on G representing X be a distance measure between x_i and x_j. Let $y = [y_1, ..., y_n]^T$ be the initial vector with value 1 or 0. If x_i is a query then y_i = 1 otherwise y_i = 0. $r : X \rightarrow R$ is a function used to determine the ranking measure r_i for every data point x_i. The cost function is given by Eq. (1).

$$O(r) = \frac{1}{2}(\sum_{i,j=1}^{n} w_{ij} \| \frac{1}{\sqrt{D_{ii}}} r_j \|^2 + \mu \sum_{i=1}^{n} \| r_i - y_i \|^2) \quad (1)$$

where, the smoothness constraint in the first term in $O(r)$ allows close ranking scores for points belonging to same neighborhood. The fitting constraint is incorporated in the second term to map the starting value given to the labels with the obtained results. μ is a parameter for regularization and a diagonal matrix D is represented as $D_{ii} = \sum_{j=1}^{n} w_{ij}$. Symmetrical normalization of W using D is given by S where $S = D^{\frac{-1}{2}} W D^{\frac{-1}{2}}$ and $\alpha = \frac{1}{1+\mu}$, we get the optimal r (cost minimization) as shown in Eq. (2).

$$r* = (I_n - \alpha S)^{-1} y \quad (2)$$

To compare the ranking generated by MR, we also consider an alternate method where ranking is computed using Euclidean Distance (ED). If the points r, s are the data points and n is the feature measure, then,

$$ED(x, y) = \sum_{i=1}^{n} \sqrt{(s_i - r_i)^2} \quad (3)$$

ED is simple as it does not involve any complex calculations and can be used to compare results with the proposed approach using the MR algorithm. Further, the MR model is further extended with a relevance feedback cycle, which is discussed in Section III.

III. Relevance Feedback

Relevance feedback is a technique which includes involvement of the user in the retrieval process so as to improve the performance ranking of the retrieval model. For providing relevance feedback in the MR model, the user is requested to provide feedback on the retrieved images indicating if they were relevant/irrelevant with reference to the query image by entering +1 or -1 for each. This feedback loop basically helps in further improving the retrieval performance and helping towards capturing the semantic concepts of the query correctly.

In MR with relevance feedback, the initial vector y is assigned as 1 for the query. However, for each of the other images i, y_i is assigned as the average of the relevance information provided in previous queries. For example, let y_0=1 signify the first image as the query. If y_0 has been queried before, there would be a relevance feedback value of +1 or -1 for the retrieved images corresponding to the query. If the image has been queried n times, the average of the n feedbacks per image is considered to form the initial vector. Also, images not retrieved have no feedback, i.e. $y = 0$. Thus, in initial vector y, an image marked relevant for a query previously has a positive weight, while an irrelevant image has negative initial weight. The improvements observed in performance due to this processed are presented in Section IV.

IV. Experimental Results

The proposed medical image retrieval model with ranking based on MR was implemented using the Matlab 2014Ra setup and the hardware architecture used was a workstation with 16GB RAM and 2.0 GHZ (*2) CPU. The results were analyzed on the IRMA (Image Retrieval in Medical Applications) dataset [13] that consists of X-ray images collected during daily routine radiological work at RWTH Aachen University. It contains approximately 12,000 labeled X-ray images of multiple organs such as spine, palm bones, knees, ankles, etc. Fig. 1 shows some sample images.

Fig. 1. Sample images from medical dataset IRMA

Each of the dataset images were preprocessed so that image features were extracted as per the defined feature set of mean, skewness, variance, kurtosis, entropy, and energy. Standard Matlab functions were used for this purpose. The neighborhood window for every pixel in the calculation of energy and entropy is maintained to be 3*3. The feature set of all images together were then represented as an $N * N$ matrix, where N denotes the total number of images. In the current experimental setup, N is of the order of 12000. The $N * N$ matrix is a representation of every data point's relation with other data points. A ranking is assigned to every data point based on the similarity with the query using the ranking function as described in section II.

The results were obtained through running a set of queries, from the categories of spinal cord, lungs and the palm. The evaluation metric used is True positive rate (TPR), i.e. precision at 10, 20 and 30 retrievals respectively. The mean average precision (MAP) is computed at each of these scales. Precision at k is defined as the amount of relevant results in the first k search results for the given query. In the considered setting, precision at k is the number of images of the same category as that of the query image in the first k retrieval results. The MAP at k is the mean of the precision at k values obtained for various query images.

$$MAP_k = 1/Q \sum_{q=1}^{q=Q} P_k(q) \tag{4}$$

where, P_k is the precision at k for query q and Q is the total count of queries.

The results for precision at each scale is tabulated in Table I and the MAP is presented in Table II. It was observed that MR achieved better precision than ED by an average of 2-10%. From Table 2, it can be observed that MR shows a considerable improvement in ranking performance as compared ED approaches. The plot of MAP vs. Precision@k is shown in Fig. 2 for the two ranking techniques, which also shows significantly better performance in the case of the MR approach.

TABLE I
Precision at k Performance with ED and MR approaches

Query Image	Metric	MR	ED
1	Precision@10	75%	70%
1	Precision@20	75%	70%
1	Precision@30	72%	70%
3	Precision@10	14%	12%
3	Precision@20	17%	10%
3	Precision@30	12%	10%
17	Precision@10	81%	80%
17	Precision@20	90%	85%
17	Precision@30	92%	86%
51	Precision@10	23%	15%
51	Precision@20	37%	20%
51	Precision@30	45%	40%

TABLE II
Mean Average Precision values at k=10,20,30 (MAP@k)

Model	MAP@10	MAP@20	MAP@30
MR	48	55.19	52.35
ED	44.25	46.25	51.5

The relevance feedback loop comes into play when the results are generated for the first image query. A simple user interface using which the user can submit a query image was developed, and the generated search results are also displayed using the same UI. For every query q in Q, the top 30 relevant images as generated by the system are displayed to the user. Once the results are generated, the user is asked for feedback on the relevance of the results generated, in the context of the particular query. To leverage this feedback, a dynamic labeling strategy is adopted. The user-indicated feedback is a binary feedback, i.e. a particular image in the result-set is either relevant (considered as +1) or irrelevant (taken as - 1) respectively. This is stored as an array of k elements where k is the number of retrieved images. If the same image is queried repeatedly (more than a specified threshold of frequency), an initial vector 'y' is updated to the average of all user responses for that query.

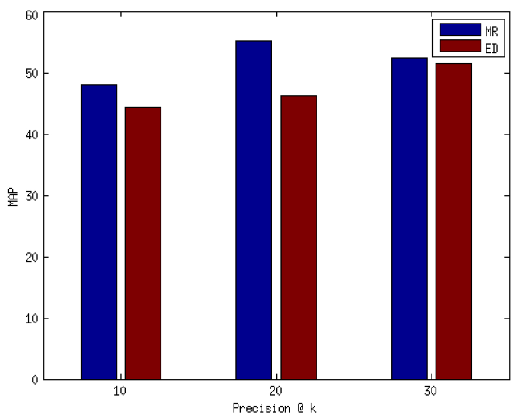

Fig. 2. MAP vs Precision@k

The difference between results retrieved before and after relevance feedback can be seen from the Fig. 3 and Fig. 4 respectively. The image in the first row is that of a spine used as a query image. The images following from the second row are the retrieval results for the query image. Spine images are retrieved at the top four positions when relevance feedback is considered as opposed to the results without feedback. Overall, it can also be observed that the number of spine images retrieved in the top 20 results in Fig.4 are more when relevance feedback is incorporated. This clearly indicates a higher relevancy result for retrievals with relevance feedback.

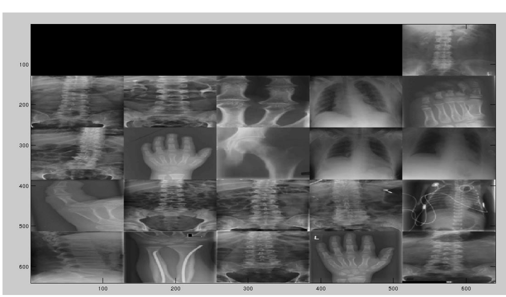

Fig. 3. Retrieval results for a spine-related image query (without relevance feedback)

V. Conclusion and Future Work

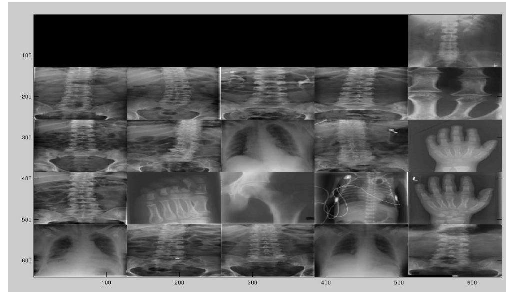

Fig. 4. Improvement in retrieval results for the same spine-related image query over time (with relevance feedback)

In this paper, Manifold Ranking technique based medical image retrieval method was presented. In addition to large-scale ranking support offered by MR for datasets, a user level relevance feedback mechanism was also incorporated, which further enhanced retrieval performance. Experimental analysis showed that MR was economical, accurate and effectively adopted to real world CBIR applications such as medical image retrieval. As future work, we intend to explore ways for understanding and integrating latent semantic concepts of medical images, to further enhance retrieval. To improve the retrieval time for large datasets, the underlying graph construction could be deployed as a parallel or distributed architecture, for supporting real-time applications like Clinical Decision Support systems.

Acknowledgement

We thank Dr. TM Deserno, Dept. of Medical Informatics, RWTH Aachen, Germany for providing us with the dataset.

References

[1] Niblack, C. W., Barber, R., Equitz, W., Flickner, M. D., Glasman, E. H., Petkovic, D., & Taubin, G. (1993, April). QBIC project: querying images by content, using color, texture, and shape. In IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology (pp. 173-187). International Society for Optics and Photonics.

[2] Lehmann, T. M., Guld, M. O., Thies, C., Fischer, B., Keysers, D., Kohnen, M., ... & Wein, B. B. (2003, May). Content-based image retrieval in medical applications for picture archiving and communication systems. In Proceedings of SPIE (Vol. 5033, pp. 109-117).

[3] El-Kwae, E. A., Xu, H., & Kabuka, M. R. (2000). Content-based retrieval in picture archiving and communication systems. Journal of Digital Imaging, 13(2), 70-81.

[4] Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schlkopf, B. (2004). Learning with local and global consistency. In Advances in neural information processing systems (pp. 321-328).

[5] Zhou, D., Weston, J., Gretton, A., Bousquet, O., & Schlkopf, B. (2004). Ranking on data manifolds. In Advances in neural information processing systems (pp. 169-176).

[6] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333-389.

[7] Lee, D. L., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space model. IEEE software, 14(2), 67-75.

[8] Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. Journal of documentation, 35(4), 285-295.

[9] He, J., Li, M., Zhang, H. J., Tong, H., & Zhang, C. (2004, October). Manifold-ranking based image retrieval. In Proceedings of the 12th annual ACM international conference on Multimedia (pp. 9-16). ACM.

[10] Shyu, C. R., Brodley, C. E., Kak, A. C., Kosaka, A., Aisen, A. M., & Broderick, L. S. (1999). ASSERT: A physician-in-the-loop content-based retrieval system for HRCT image databases. Computer Vision and Image Understanding, 75(1-2), 111-132.
[11] Keysers, D., Ney, H., Wein, B. B., & Lehmann, T. M. (2003). Statistical framework for model-based image retrieval in medical applications. Journal of Electronic Imaging, 12(1), 59-68.
[12] Deselaers, T., Keysers, D., & Ney, H. (2004, September). FIRE-flexible image retrieval engine: ImageCLEF 2004 evaluation. In CLEF (pp. 688-698).
[13] Lehmann, T.M., Gold, M.O., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H. and Wein, B.B., 2004. Content-based image retrieval in medical applications. Methods of information in medicine, 43(4), pp.354-361.