AUTOMATED FEEDBACK EXTRACTION FOR MEDICAL IMAGING RETRIEVAL

Weidong Cai[1], *Fan Zhang*[1], *Yang Song*[1], *Sidong Liu*[1], *Lingfeng Wen*[2], *Stefan Eberl*[2], *Michael Fulham*[2,3], *Dagan Feng*[1]

[1]BMIT Research Group, School of IT, University of Sydney, Australia
[2]Department of PET and Nuclear Medicine, Royal Prince Alfred Hospital, Australia
[3]Sydney Medical School, University of Sydney, Australia

ABSTRACT

Content-based image retrieval (CBIR) has been widely used in many medical applications by providing objective depictions and the initial screening to facilitate the manual interpretations by the radiologists. To achieve accurate retrieval results, relevance feedback is usually incorporated into CBIR to refine the retrieved items, but its effectiveness is restricted by the huge number of medical cases. Therefore, in this study we propose an automated feedback extraction method to exclude the involvement of radiologists. Instead of incorporating the feedbacks from them, the similarity relationship between the initial retrieval results and all candidate images is used to indicate the preferences of these retrieved items regarding to the query, *i.e.*, relevance or irrelevance, and to further re-rank the candidates. The experimental results on a publicly available brain image dataset for neurodegenerative disorder diagnosis demonstrate the promising retrieval performance of the proposed method.

Index Terms— *medical imaging; CBIR; automated relevance feedback*

1. INTRODUCTION

Content-based image retrieval (CBIR) plays an important role in establishing the quantitative correlations among images by analyzing their content, instead of searching by keywords or tags. CBIR systems offer the great opportunity for us to access the large collections of pre-diagnosed medical cases, thus provide important insights for understanding the disease pathology [1].

The general work flow of CBIR systems consists of the following steps [2, 3]: feature extraction, similarity calculation, and relevance feedback. While the first two steps form a basic retrieval process, relevance feedback refines the retrieval results if the top-ranked items are not fully satisfactory. Current works on content-based medical image retrieval mainly focus on the first two steps [4-9], suggesting that the more advanced feature designs could deliver better retrieval performances. However, even with the most advanced features, it remains challenging to retrieve the images accurately, since the systems only have one chance to make the prediction [10]. Relevance feedback would be helpful by further revising the retrieval results.

Relevance feedback is based on users' preferences upon the initial retrieval results. These feedbacks are further repeatedly used to predict the relevance of the candidate images in the database to the queries. The common form of such interaction is to ask the users to manually provide feedbacks on the relevance of the retrieved items [11]. A number of relevance feedback techniques have been proposed for medical imaging retrieval, such as query point movement [12] and feature re-weighting [13].

Although the initial retrieval stage has already eliminated some irrelevant items and left a relatively small number of cases for tagging, it is still too much of a burden for the radiologists to provide feedbacks. In addition, the interpretations of different radiologists would be varied, and thus the final retrieval results would be biased by the individual perception of subjectivity.

Instead of using the feedbacks provided by the users, the relationship among images themselves provides important clues for telling the relevance/irrelevance of the retrieval results, and thus can be used to assist the retrieval. For example, a ranking scheme for dementia diagnosis (distinguishing dementia patients from normal subjects) using both similar and dissimilar items was proposed in [14]. Although it shows the correlations among images are beneficial for image retrieval and can be used as a form of relevance feedback, it is hard to incorporate this scheme for multi-class retrieval. In addition, we believe that the relationship among images can be further utilized, not merely extracting similar and dissimilar items.

In light of above, we would like to design an automated feedback extraction method to assist content-based medical image retrieval. The contributions of the proposed method are twofold: first, the similarity relationship between the retrieved items and candidates is used to infer the feedbacks, *i.e.*, the preferences of the retrieved items and the relevance of the candidates to the query; second, an iterative ranking calculation method is designed to update weights of the retrieved items and the candidates to quantize the above feedbacks. The proposed method focuses on relevance

feedback design and is based on few assumptions about the problem domain, therefore it is generally applicable to other medical or general imaging domains. In this study, we focus on brain image analysis to demonstrate our method. Evaluated on a publicly available brain imaging set, our method shows a marked improvement over the traditional CBIR methods with no relevance feedback.

2. METHOD

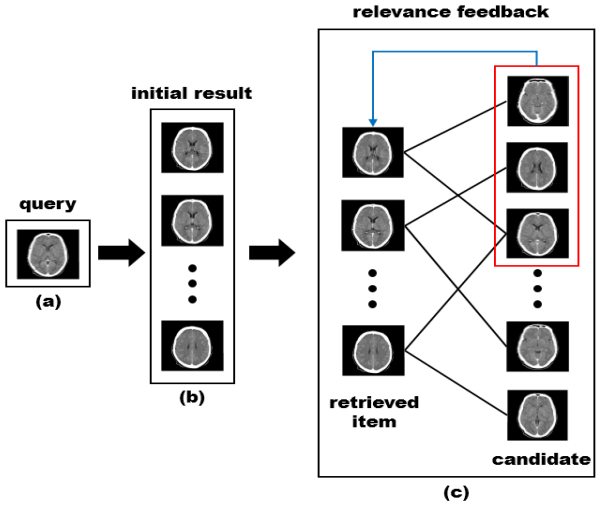

Fig. 1. Illustration of the CBIR with proposed automated feedback extraction for sample brain images: (a) the query image, (b) the initial retrieval list, and (c) the iterative feedback extraction procedure. The retrieved items and candidates are ranked according to the preference scores and the relativity scores respectively, which are calculated based on the similarities among these images (shown as the black lines between retrieved items and candidates). The top-ranked candidates (shown in the red rectangle) are selected as the retrieved items iteratively (shown with the blue arrow) to further re-rank all candidates.

The overall procedure of the proposed method is illustrated in Fig. 1. For a given query image, a basic retrieval process is conducted at the first stage, which produces a ranked list of initial retrieval results based on the visual feature extraction and similarity calculation (introduced in Section 2.1). During the relevance feedback stage, the similarity relationship between the retrieved items and all candidates is established and further utilized to re-rank the candidates iteratively (introduced in Section 2.2).

Formally, for the query image I_q, the initial retrieval list is denoted as $R_{ini}(I_q)=\{I_r:r=1\ldots M\}$ with the first M ranked images from all candidates in dataset $D=\{I_c:c=1\ldots N\}$ with total N images ($M \ll N$). During each iteration $t \in \{1\ldots T\}$ of total T iterations, the relativity score $rel(I_c)$ of candidate I_c and the preference score $pref(I_r)$ of retrieved item I_r are calculated based on the similarity measures, *i.e.*, $SIM(R_t(I_q), D)$, between all candidates from D and all retrieved items from $R_t(I_q)=\{I_r:r=1\ldots M\}$. For iteration t, $R_t(I_q)$ was achieved by selecting the top-ranked M candidates at iteration (t-1). The final retrieval results for I_q, *i.e.*, $R_{fin}(I_q)=\{I_k:k=1\ldots K\}$ are obtained by selecting the first K images from $R_T(I_q)$ ($K \leq M$).

2.1. Features Extraction and Similarity Calculation

At the first stage, features are extracted from the query image I_q and all candidates to describe the images numerically. The similarities between them are computed to derive the initial retrieval results $R_{ini}(I_q)$ with the basic retrieval process. Since the objective of this work is relevance feedback design, without over-emphasizing on the feature design, we simply extract the low level features (introduced in Section 3.1), denoted by $feat(I)$. The similarities between the query image and candidates are calculated with the Euclidean metric. The top M candidates are selected as the $R_{ini}(I_q)$, which are also used as the first retrieved items $R_1(I_q)$ in the relevance feedback stage.

2.2. Relevance Feedback

At the second stage, we aim at obtaining a more precise retrieval result with relevance feedback. Different from the manual feedback approach, the proposed automated scheme does not depend on the involvement of users, and the interactions between the retrieved items and candidates are used to extract the feedbacks. It contains two components: *preference indication* and *relativity updating*. As for the first component, the preferences upon the retrieved items regarding to the query, *i.e.*, relevance and irrelevance, are inferred based on whether their neighbouring candidates are related to the query, instead of provided by the users. The preference is represented quantitatively as the preference score, $pref(I_r)$ for the retrieved item I_r, with higher $pref$ indicating more preferred. As for the second component, the candidates are ranked according to the preferences of the neighbouring retrieved items. The ranking score of candidate image I_c is denoted as the relevance score $rel(I_c)$, with higher rel indicating more relevance.

These two components contribute to each other reciprocally. Regarding the query image I_q, the relativity score $rel(I_c)$ of candidate I_c would be high if it is similar to the highly preferred retrieved item I_r, and the preference score $pref(I_r)$ of I_r would be high if it is close to the more relevant candidate I_c. The relativity score of I_c is formulated as the sum of preference scores of its neighbouring retrieved items, as shown in Eq. (1), and the same to preference score of I_r, as shown in Eq. (2). $\overrightarrow{rel}$ and $\overrightarrow{pref}$ are the vectors of relativity scores of retrieved items in $R(I_q)$ and preference score of candidate images in D,

$$\overrightarrow{rel} = A \bullet \overrightarrow{pref} \tag{1}$$

$$\overrightarrow{pref} = A^T \bullet \overrightarrow{rel} \tag{2}$$

where A is a $N \times M$ matrix indicating the neighborhood between retrieved items and candidates, *i.e.*, $A(n,m)=1$ if $I_c(n)$ is the neighbour of $I_r(m)$; otherwise, $A(n,m)=0$.

Specifically, at each iteration $t \in \{1 \ldots T\}$, given the retrieved items $R_t(I_q)$ and the candidates D, a ranking calculation is designed to maintain and update the relevance scores and preference scores according to Eq. (1) and Eq. (2). At the beginning, adjacency matrix A is derived by parsing the similarity network $SIM(R_t(I_q), D)$, whose elements are the Euclidean distance measures between the feature vectors $feat(I_r)$ and $feat(I_c)$. For each retrieved image I_r, its first S most similar candidates are selected as neighbours to set $A(I_c, I_r) = 1$.

Then, Eq. (1) and Eq. (2) are solved with an iterative algorithm, which makes use of the relationship between the retrieved items and candidates, *i.e.*, A. The preference score vector $\overrightarrow{pref}$ is initially set as $(1 \ldots 1)^T \in \boldsymbol{R}^M$, indicating that the initial preferences upon the retrieved items are equivalent. Next, the relativity score of candidate I_c is computed by:

$$rel(I_c) = \sum_{A(I_c, I_r)=1} pref(I_r) \tag{3}$$

Thus, for I_c, its relativity depends on the preferences of retrieved items that are connected to it. Following that, the preference score of retrieved item I_r is updated by:

$$pref(I_r) = \sum_{A^T(I_r, I_c)=1} rel(I_c) \tag{4}$$

Hence, for I_r, its preference is re-ranked since the relativities of its neighbouring candidates have changed.

The above two steps are conducted iteratively to update the preference score and relativity score vectors. At the beginning of each iteration t, the retrieved items are selected from the top-ranked candidates in iteration (t-1); at the end of the iteration, both of these vectors are normalized so their squares sum to 1: $\sum_{c \in [1,N]} (\overrightarrow{rel}(c))^2 = 1$ and $\sum_{r \in [1,M]} (\overrightarrow{pref}(r))^2 = 1$. We found in the experiments that 20 iterations are sufficient to derive stable vectors and generate a good performance.

3. EXPERIMENT AND EVALUATION

3.1. Dataset and Evaluation

The proposed method was evaluated on the dataset obtained from the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database [15]. A total of 331 subjects were randomly selected, with 77 cognitive normal subjects, 169 mild cognitive impairment (MCI) patients, and 85 Alzheimer's disease (AD) patients, annotated with CN, MCI and AD respectively.

In the experiments, 7 different types of features were extracted from the multi-modal brain imaging data to make sure that the proposed method can work in various circumstances and is independent of the feature design. The Mean Index [4], Fisher Index [5], and DOG Features (DOG Area, DOG Contrast, DOG Mean) [6] were extracted from positron emission tomography (PET) data, and the Solidity [7] and Volume [8] features were extracted from MRI data.

The leave-one-out cross-validation across the whole dataset was adopted for performance evaluation. The performances were quantitatively measured by the average precision of the queries, as:

$$AP = \frac{\sum_{I_q \in D} (\sum_{I_k \in R_{fin}(I_q)} acc_{I_q}(I_k) / |R_{fin}(I_q)|)}{|D|} \tag{5}$$

where $|\blacksquare|$ is the size of image set, and $acc_{I_q}(I_k)$ is the accuracy that assigning I_k with the category with I_q. Since MCI usually represents the transition state from CN to AD, an elastic relevance criterion is incorporated for computing *acc*, as introduced in [16]. The values of *acc* of I_k in terms of I_q are shown in Table I.

TABLE I
THE ELASTIC RELEVANCE CRITERION TABLE

$I_q \setminus I_k$	CN	MCI	AD
CN	1	0.25	0
MCI	0.25	1	0.25
AD	0	0.25	1

3.2. Results

To illustrate the advantages of iterative ranking sore updating, *i.e.*, preference score and relativity score, and to show the responses of the proposed method to the different numbers of final outputs, Fig. 2 and Fig. 3 show the distributions of *AP* given various numbers of iterations T (from 0 to 19) and different numbers of final retrieval results K (from 1 to 9) on the combination of the extracted 7 features. Except for these two parameters, the proposed method also introduces the number of retrieved items M, and the number of neighbours S for constructing matrix A, therefore the average *AP* was computed from 5 to 50 for both of M and S.

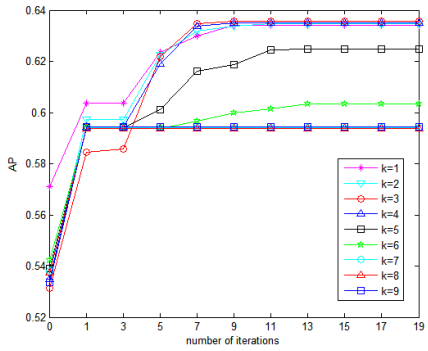

Fig.2. Effect of the number of iterations on the combined 7 features.

Fig. 2 shows the effect of the iterative updating. It follows two main patterns regarding to various numbers of final retrieval results K. The first one comes from the smaller k, *i.e.*, from 1 to 5. As the number of iterations increases, the average *AP*s keep climbing gradually and stay constant after around 10 iterations. The other is from the larger k, *i.e.*, from 6 to 9. The average *AP* distributions remain stable after the first iteration. The reason is that: for the smaller size of

outputs, the proposed method continuously updates the ranking scores so that the few most similar items can rank at the top among the iteration-1 retrieved items; for the larger size, the similar items have already been incorporated into the results after the iteration-1, and further updating would not impact on the performance since the items in the final retrieval results contribute equally to *AP*. Although the performances are diverse, the *AP*s from both patterns are higher than the initial retrieved result (indicated with iteration-0) from the basic retrieval. Therefore, it shows that the retrieval results are improved by updating the retrieved items and the candidates.

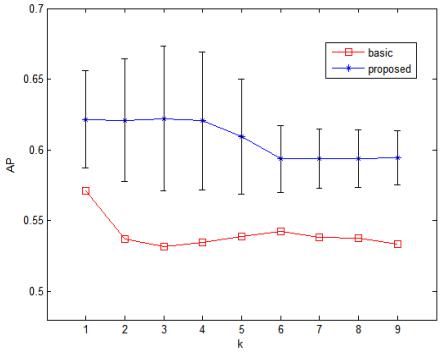

Fig.3. Effect of the number of retrieval results on the combined 7 features.

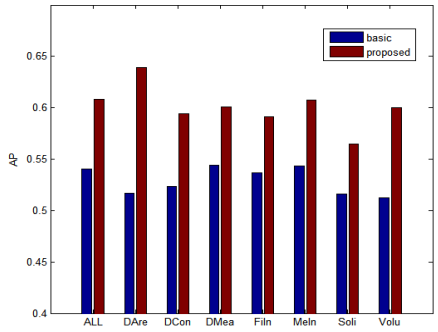

Fig. 4. The comparisons of the average *AP* between the basic retrieval and proposed method for each individual feature and the combination of them.

As for the performances of the proposed method regarding to the different numbers of final retrieval results, the comparisons with the basic retrieval are shown in Fig. 3, with standard deviation upon the average *AP* across all *M* and *S*. The best enhancements come from the lower number of final retrieval results. A best *AP* of nearly 68% (adjusted by the standard deviation) was achieved by the proposed method when k=3, nearly 15% higher than the basic approach. In addition, as k is increasing, the standard deviation becomes relatively smaller, which suggests that the performance tends to be stable if more retrieval results are retrieved. Overall, the higher *AP*s indicate that the proposed method outperforms the basic retrieval approach.

Finally, the average *AP*s across all k for the combination of the 7 descriptors as well as each of them are shown in Fig. 4. The proposed method achieves better retrieval results for all of the given features. The best enhancement with an over 10% *AP* increase can be found when using the DOG Area features.

4. CONCLUSIONS

In this study, we propose an automated feedback extraction method to improve the performance of CBIR for medical imaging analysis. Different from the common manual relevance feedback techniques, the proposed method does not depend on preference indications from the uses, but relies on the feedbacks extracted from the similarity relationship between retrieved results and candidates. Our method is evaluated using the ADNI database for AD and MCI retrieval, and the results show its promising performance. Since it is independent of the features design, we suggest that the proposed method is broadly applicable to other medical or general image retrieval domains.

5. REFRENCES

[1] H. Müller, *et al.*, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *International journal of medical informatics*, vol. 73, pp. 1-23, 2004.

[2] W. Cai, J. Kim, D. Feng, "chapter 4 content-based medical image retrieval," in *Biomedical Information Technology*, ed: Elsevier, 2008.

[3] W. Cai, D. Feng, R. Fulton, "Content-Based Retrieval of Dynamic PET Functional Images", I*EEE Trans. Information Technology in Biomedicine*, Vol.4, No.2, pp152-158, 2000.

[4] W. Cai, S. Liu, *et al.*, "3D neurological image retrieval with localized pathology-centric CMRGlc patterns," in *ICIP*, 2010, pp. 3201-3204.

[5] S. Liu, *et al.*, "Generalized regional disorder-sensitive-weighting scheme for 3D neuroimaging retrieval,"in *EMBC*, 2011,pp.7009-7012.

[6] M. Toews, W. Wells III, D. L. Collins, *et al.*, "Feature-based morphometry: Discovering group-related anatomical patterns," *NeuroImage*, vol. 49, pp. 2318-2327, 2010.

[7] P. G. Batchelor, *et al.*, "Measures of folding applied to the development of the human fetal brain," *IEEE Trans. Medical Imaging*, vol. 21, pp. 953-965, 2002.

[8] R. A. Heckemann, S. Keihaninejad, *et al.*, "Automatic morphometry in Alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 56, pp. 2024-2037, 2011.

[9] S. Liu, et al, "Multifold Bayesian Kernelization in Alzheimer's Diagnosis", in *MICCAI*, 2013, pp303-310.

[10] S. Liu, et al, "A Supervised Multiview Spectral Embedding Method for Neuroimaging Classification", in *ICIP*, 2013, pp602-605.

[11] X.S. Zhou, *et al.*,"Relevance feedback in image retrieval:A comprehensive review," *Multimedia Sys.*, vol. 8, pp. 536-544, 2003.

[12] C. Akgül, D. Rubin, S. Napel, C. Beaulieu, *et al.*, "Content-Based Image Retrieval in Radiology: Current Status and Future Directions," *Journal of Digital Imaging*, vol. 24, pp. 208-222, 2011.

[13] M. M. Rahman, P. Bhattacharya, and B. C. Desai, "A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback," *IEEE Trans. Information Technology in Biomedicine*, vol. 11, pp. 58-69, 2007.

[14] D. Unay and A. Ekin, "Dementia diagnosis using similar and dissimilar retrieval items," in *ISBI*, 2011, pp. 1889-1892.

[15] W. J. Jagust, *et al.*, "The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core," *Alzheimer's & Dementia*, vol. 6, pp. 221-229, 2010.

[16] S. Liu, W. Cai, *et al.*, "Multi-channel Brain Atrophy Pattern Analysis in Neuroimaging Retrieval," in *ISBI*, pp. 206-209, 2013.