I.J. Information Engineering and Electronic Business, **2011, 3, 16-22**
Published Online June 2011 in MECS (http://www.mecs-press.org/)

A Domain Knowledge Based Approach for Medical Image Retrieval

Haiwei Pan
College of Computer Science and Technology
Harbin Engineering University Harbin, China
heaven_007cn@yahoo.com.cn

Xiaolei Tan Qilong Han Guisheng Yin
College of Computer Science and Technology
Harbin Engineering University Harbin, China
{tanxiaolei, hanqilong, yinguisheng}@hrbeu.edu.cn

Abstract: The high incidence of brain disease, especially brain tumor, has increased significantly in recent years. It is becoming more and more concernful to discover knowledge through mining medical brain image to aid doctors' diagnosis. Image mining is the important branch of data mining. It is more than just an extension of data mining to image domain but an interdisciplinary endeavor. Image clustering and similarity retrieval are two basilic parts of image mining. In this paper, we introduce a notion of image sequence similarity patterns (ISSP) for medical image database. ISSP refer to the longest similar and continuous sub-patterns hidden in two objects each of which contains an image sequence. These patterns are significant in medical images because the similarity for two medical images is not important, but rather, it is the similarity of objects each of which has an image sequence that is meaningful. We design the new algorithms with the guidance of the domain knowledge to discover the possible Space-Occupying Lesion (PSO) in brain images and ISSP for similarity retrieval. Our experiments demonstrate that the results of similarity retrieval are meaningful and interesting to medical doctors.

Index Terms-Data mining; image mining; similarity retrieval; domain knowledge

1 INTRODUCTION

Advances in image acquisition and storage technology have led to tremendous growth in very large and detailed image databases [1]. A vast amount of image data is generated in our daily life and each field, such as medical image (CT images, ECT images and MR images etc), satellite images and all kinds of digital photographs. These images involve a great number of useful and implicit information that is difficult for users to discover.

Image mining can automatically discover these implicit information and patterns from the high volume of images and is rapidly gaining attention in the field of data mining. Image mining is more than just an extension of data mining to image domain. It is an interdisciplinary endeavor that draws upon computer vision, image processing, image retrieval, machine learning, artificial intelligence, database and data mining, etc. While some of individual fields in themselves may be quite matured, image mining, to date, is just a growing research focus and is still at an experimental stage. Research in image mining can be broadly classified to two main directions: (1) domain-specific applications; (2) general applications [2]. The focus in the first direction is to extract the most relevant image features into a form suitable for data mining [4,8,9] and the latter is to generate image patterns that may be helpful in understanding of the interaction between high level human perceptions of image and low level image features [1,10]. Data mining in medical images belongs to the first direction.

Brain tissue is human's advanced nerve center, so its function is particularly important. The disease affecting the brain has received much attention in the domain of medicine. In China, about 40,000 to 60,000 persons suffer from brain tumor every year and about 16% of these persons are children. Especially during these years, the incidence of brain disease (especially brain tumor) has increased significantly and the quality of human's living even their lives has been endangered greatly. Therefore, the early diagnosis of brain diseases is becoming more and more crucial and is directly working on patients' treatment. That is why data mining in medical images for assisting medical staff is so significant. Furthermore, it is a greater challenge because of referring to the special domain.

Computerized Tomography (CT) is one of the most important techniques that are used to diagnose by medical doctors. Brain CT scan of each patient (as an object below) is an image sequence in which each one is an image of a layer every a few millimeters from calvaria. There exists a certain spatial relationship between images in the sequence. We will try to discover knowledge from this kind of image dataset by means of data mining technique.

This paper presents a new method to retrieve similar objects each of which includes an image sequence. The novelty includes three directions. The first is to make use of medical domain knowledge efficiently to guide data mining. The second is that we utilize two different clustering algorithms on pixels to discover the possible brain diseases that are called Space-Occupying Lesion

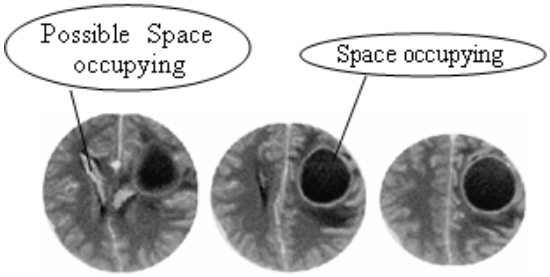
Copyright © 2011 MECS

(SO) by doctors, as shown in figure 1(c). The third is that we introduce a notion of image sequence similarity patterns ISSP for similarity retrieval.

The rest of the paper is organized as follows: section 2 is the statement of problem. Pre-processing is presented in section 3 and Detecting PSO Based on Pixel's Clustering is introduced in section 4. Section 5 presents Similarity Retrieval Based On ISSP. Conclusions and future research are presented in section 6.

(a) Normal object's brain image before preprocessing

(b) Normal object's brain image after preprocessing



(c) An abnormal object's brain image sequence after preprocessing, further from calvaria from the first to the last

Figure 1: An example of normal and abnormal brain image

2 STATEMENT OF PROBLEM

At present, the main work of data mining on medical images [3,4,5,6,7,12,13] has two characteristics: (1) research content is the images in the medical image database, not the objects with medical images. For example, it is possible to classify images in the same object into the different class because they always have different morphological SO. This determines the type of knowledge that will be mined; (2) research method is to extract features from images to form feature attributes and use data mining on these attributes, not to consider the fundamental element – pixel's significance. In fact, medical doctors make a diagnosis mainly according to medical knowledge and the tone of pixels. Figure 2 shows the framework of these precious works, where IM_i is an image and F_{ij} is a feature in IM_i. Also, these work paid little attention to the guidance effect of domain knowledge to data mining.

Our research content is objects each of which contains a series of images and the images from different objects maybe have different intensity, see figure 3. The images

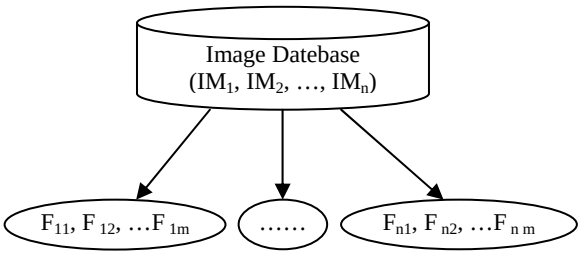


Figure 2: Framework of Precious Works. IM_n is an image and $F_{n\,m}$ is one of the features extracted from IM_n.

from one object have the same structure but come from the different layers of a brain. They have similar pixel density distribution and it is possible for two spatial-adjacent images to be similar in one object, see figure 1(c). Each image contains some PSOs that also have a spatial relationship in an image. Therefore, the objects with the similar image sequence patterns (ISP) should have similar clinical manifestations. It is very helpful to assist medical doctors to make a diagnosis.

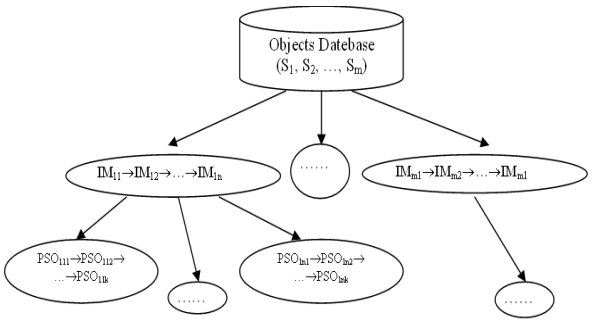


Figure 3: Framework of Our Work

3 PRE-PROCESSING

Since the images we got were raw CT scans that were scanned at different illumination conditions, some of them appeared too bright and some were too dark. We should digitize them to no loss, no compression and 256 gray scale images through special medical scanner. As brain CT scan of each object comprises several or more than ten images and there is a certain spatial relationship between these images, this greatly increases the quantity and complexity of data that is to be processed.

In preprocessing, our work uses domain knowledge effectively to remove the noisy data. A brain CT image mainly consists of three parts: noisy data, skull and cerebrum. The noisy data includes the black background and some additive information on it, such as CT identification, date and patient's name etc. This information is not only helpless but revealing patient's privacy. We are only interested in cerebrum. So we use domain knowledge (for short DK1) and cropping technique of image processing to gain it.

DK1 --- human's brain skull has the highest density and surrounds the cerebrum.

Copyright © 2011 MECS

Table 1: Each object is formed as a record in the table after its images are preprocessed.

ID	IM
001	001.01→001.02→001.03→…→001.07
002	002.01→002.02→002.03→…→002.13
…	……
n	n.01→n.02→n.03→…→n.10

That is, the skull is a cricoid area in the image with the whitest pixels that separate cerebrum from the noisy data, see figure 1(a). It becomes easy to remove the noisy data by using cropping technique and keep the interesting region with the guidance of DK1. After image preprocessing, all the objects are formed as follows, see table 1. Each object has a unique identification (ID) and its image part (IM) is a preprocessed image sequence where every code is composed of id, a dot and the image sequence number, and the arrow represents the spatial relationship.

4 DETECTING POSSIBLE SO (PSO) BASED ON PIXEL'S CLUSTERING

The following domain knowledge (for short DK2) is used to direct clustering algorithm.

(1) The normal persons have nearly the same brain structure that is evident to be lateral symmetry. That is, the distribution of density in the left hemisphere of the brain is almost identical with the right, see figure 1 (b). If there is SO in either side, its density will change and destroy the symmetry;

(2) If one object has SO, it is more possible that this SO will be shown in some continuous images, see figure 1 (c).

4.1 Basic Definitions

Let $S=\{S_i \mid i=1\ldots m\}$ be object set; Let $S_i=\{IM_{i1}, IM_{i2}, \ldots, IM_{in}\}$ be object or ordered image set, where

(1) $IM_{i1}$and IM_{in} are the nearest and farthest image from calvaria;

(2) For any IM_{i1}, IM_{i2}, …, IM_{ip}, IM_{ip} must be the farthest image from calvaria;

For any preprocessed image IM_p, it is halved by brain midline (shown as figure 1(b)) and is composed of two parts: $IM_p(L)$ and $IM_p(R)$ present the left and right hemisphere image respectively.

For any IM_p and IM_j, they are adjacent if

(1) $IM_p \in S_i$ and $IM_j \in S_i$;

(2) p=j+1 or p=j-1;

Definition 1. Pixel set of IM_p is defined as $P=\{p_i \mid p_i$ is the pixel with coordinate (x_i, y_i) in the image $IM_p\}$, P(L) and P(R) are pixel set of $IM_p(L)$ and $IM_p(R)$ respectively.

According to the symmetry of the brain structure in DK2, we assume that the number of pixels in $IM_p(L)$ and $IM_p(R)$ are equal. That is, $|P(L)|=|P(R)| = |P|/2$.For any $p_{li} \in P(L)$, $p_{rj} \in P(R)$, they are symmetric pixel if the line between p_{li} and p_{rj} is halved vertically by brain midline. They are denoted as p_{li} and p_{ri} below.

Definition 2. We partition all pixels in pixel set P into m blocks. Pixels in the same block have the same grey level and pixels in the different blocks have the different grey level. Let $G(P)=\{g_1, g_2, \ldots, g_m\}$ be P's grey-scale (GS) set if G(P) is an ascending sort set of g_1', g_2', …, g_m' and g_i' is grey level of pixels in the i^{th} block, where g_i (i=1,…,m) is the i^{th} GS, g_1' and g_m' are P's minimum and maximum GS respectively. The GS of pixel p_i is denoted as $g(p_i)$. we call $g_{mean}(P)$ the mean GS if

$$g_{mean}(P) = \sum_{i=1}^{|P|} g(p_i)/|P|$$

Definition 3. For any P and distance function $DisA=|g_k - g_{mean}(P)|$, mid-value GS is a middle value in the GS set that minimizes DisA. Mid-value GS set is a set of mid-value GS.

Theorem 1 Mid-value GS set includes not more than two values.

Proof The value of DisA has two possibilities:

(1) If $g_k - g_{mean}(P) <0$, then $DisA= g_{mean}(P) - g_k$;

(2) If $g_k - g_{mean}(P) \geq 0$, then $DisA= g_k - g_{mean}(P)$;

In the first case, if there exist more than two GS that make the value of DisA be a certain minimum ϒ: g_1', g_2', … , g_k'(k>2), then they must satisfy the equation $g_1'=g_2'= \ldots =g_k'$. This doesn't agree with the definition of GS set. Therefore, it is only one GS that minimizes DisA in the first case, denoted as g_{mida}.

Case 2 is proved similarly and thus it is also only one GS that minimizes DisA, denoted as g_{midb}.

If $g_{mean}(P) - g_{mida} = g_{midb} - g_{mean}(P)$, then mid-value set includes two elements. Otherwise, it only includes one of g_{mida} and g_{midb} that minimizes DisA.

For any P, if

(1) Mid-value GS set includes one element g_{mid}, and g_s is the minimum value between g_{mean} and g_{mid};

(2) Mid GS set includes two elements, g_s is the minimum value between these two values;

g_s is called Benchmark GS and another one is denoted as g_s'.

Definition 4. For pixel set P, let

$g^{(l)}=\{g_i \mid g_1 \leq g_i \leq g_1+|g_1-g_s|/2\}$ be low bound GS;

$g^{(h)}=\{g_i \mid g_m-|g_m-g_s'|/2 \leq g_i \leq g_m\}$ be high bound GS;

$g^{(b)}= g^{(l)} \cup g^{(h)}$ be bound GS;

For pixel set P, let

$P^{(l)}=\{p_i \mid g(p_i) \in g^{(l)}\}$ be low bound pixel set;

$P^{(h)}=\{p_i \mid g(p_i) \in g^{(h)}\}$ be high bound pixel set;

$P^{(b)}= P^{(l)} \cup P^{(h)}$ be bound pixel set and pixels in $P^{(b)}$ are bound pixels.

Definition 5. $\Delta g(P)$ is IM_p's difference set if for any symmetrical pixel p_{li} and p_{ri} in P, $\Delta g(P)=\{ \Delta g_i \mid \Delta g_i=g(p_{li})-g(p_{ri}), i=1,2,\ldots,|p|/2 \}$.

Definition 6. For any image sequence $<IM_{ij}, \ldots, IM_{ik}>$, if

(1) only the first and last image IM_{ij} and IM_{ik} have one adjacent IM;

(2) other IM_{ip} (if existed) has two adjacent IM;

we called that it has the property of continuity. Image sequence $<IM_{ij}, \ldots, IM_{ik}>$ has the property of discontinuity if it can't satisfy the property of continuity.

Definition 7. For any pixel p_i and a certain integer ε, the assemble of pixels is called *ε-adjacent area* (ε-AA) if distance between p_i and pixels in the assemble is not more than ε. P_i is *core pixel* (c-pixel) if its ε-AA involves no less than MP pixels that satisfy some

Copyright © 2011 MECS

conditions. P_i is *immediate density reachable* from p_j if p_i is in the ε-AA of p_j and p_j is c-pixel. P_1 and p_k are *density reachable* if for some pixels $p_1, p_2, \ldots, p_k$, any pixel p_{i+1} is immediate density reachable from p_i. P_i and p_j are *density connective* if there exists pixel p_k that is density reachable from not only p_i but also p_j.

4.2 Clustering Algorithm with the Guidance of DK2

It is very crucial step for medical doctors to determine whether there is a Space-Occupying Lesion or not in the brain images. In this paper, we use clustering method on the pixels of images to detect the PSO. Firstly, for any IM_p, we compute to get the IM_p's difference set Δg(P), then sort the absolute value of each element in Δg(P) to yield the set Δg'(P)={| Δg_i| |$\Delta g_i \in \Delta g(P)$ and for any |Δg_i|, it must be maximal in the former i elements}. Each element of Δg'(P) is regarded as an atomic cluster and hierarchical clustering from bottom to top will not stop in the light of the following similarity function of difference between pixels' GS until the number of clusters is equal to a specified value k.

similarity(C_i,C_j) = min|T_i-T_j|, where T_i and T_j are mean value of all |Δg_i| in cluster C_i and C_j respectively. The algorithm is as follows:

```
Clustering algorithm I:
Input: the set Δg'(P) and the number
of clusters k
Output: k clusters that satisfy the
similarity function similarity(Ci,Cj)
1. Each element of Δg'(P) is regarded
as an atomic cluster and compute |Ti-
Tj| of the adjacent clusters;
2. Clustering in terms of
similarity(Ci,Cj);
3. While (the number of clusters is
not equal to k) {
4. Compute |Ti-Tj| of the adjacent
clusters;
5. Clustering in terms of
similarity(Ci,Cj);}
```

According to "If there is SO in either side, its density will change and destroy the symmetry" in DK2, we can deduce that if there exists SO in the image, then pixels' GS of SO should change and the value of the elements corresponding to these pixels in Δg'(P) will be much greater than zero. Otherwise, the value of these corresponding elements in Δg'(P) will approximate zero. Therefore, we set the number of clusters to 2 as the termination condition. The first step of Clustering algorithm I scans |P|/2 elements in the set Δg'(P) one time and time complexity is O(|P|/2). The second step is to select the minimum and time complexity is O(|P|/2), too. In the third step, the loop times is related to the speed of clustering. The worst case, only two clusters is clustered to a bigger cluster in each loop. The number of time is |P|/2-3 and time complexity of the 4th and 5th step is O(|P|/2-i), where i is the i^{th} loop. Accordingly, the 3rd and 5th step for the worst case need (n-3)(n+2)/2 operations and time complexity is $O(n^2)$.

According to the symmetry of the brain structure, we single out the cluster with the greater |Δg_i| from clustering algorithm I (denoted as high difference cluster) to be the main study objects of the next step. It means that data size to be processed may be reduced. For Each Δg_i, there are two corresponding symmetric pixels p_{li} and p_{ri}. All symmetric pixels $g(p_{li})$ and $g(p_{ri})$ in the high difference cluster are judged to be whether bound GS or not, then bound pixel set is generated.

Next, the based-on density clustering method is utilized to re-cluster these bound pixel set and determine the location and size of SO in each brain image. The algorithm is as follows:

```
Clustering algorithm II:
Input: all bound pixel set, ε and MP
Output: k clusters
1. Assume that the pixel count of any
bound pixel set is bn and examine ε-AA
of these bn pixels;
2. If (ε-AA of pi involves more than
MP bound pixels)
3. Then mark pi as the c-pixel;
4. While (all c-pixels) {
Clustering all density reachable
pixels; }
```

For ε and MP in clustering algorithm II, we specify their value through learning on the normal object's brain images. The learning process is as follows: (1) run the first clustering on the normal object's IM_p and achieve the bound pixel set; (2) count (not clustering) on the bound pixel sets which, in fact, are noisy data but not SO, and compute the maximum of the radius and the count of all bound pixel set which are the greatest lower bound of ε and MP. Time complexity of this algorithm is O(nlogn). The k clusters generated from this algorithm are k PSOs.

5 SIMILARITY RETRIEVAL BASED ON ISSP

In this section, we will: (1) discover the image sequence pattern (ISP) of one object; (2) discover the image sequence similarity patterns (ISSP) of two objects.

5.1 Mining ISP of one object

The whole PSO in an image can be found using the above algorithm. For these PSOs, we denote each PSO as the following: <H(L), (x_i, y_i), (x_{a1}, x_{a2}), (y_{b1}, y_{b2})>, where H (L) represents the high (low) bound GS, (x_i, y_i) is the coordinate of the center of this PSO, (x_{a1}, x_{a2}) and (y_{b1}, y_{b2}) are the max and min x and y coordinate of the PSO. They are computed by the following formula:

$$x_i = \frac{1}{k}\sum_{j=1}^{k} x_j \quad (1) \qquad y_i = \frac{1}{k}\sum_{j=1}^{k} y_j \quad (2)$$

$$x_{a1} = \max_{j=1}^{k}(x_j) \quad (3) \qquad x_{a2} = \min_{j=1}^{k}(x_{ij}) \quad (4)$$

$$y_{b1} = \max_{j=1}^{k}(y_j) \quad (5) \qquad y_{b2} = \min_{j=1}^{k}(y_j) \quad (6)$$

We take the center of the brain as the origin of coordinates. If $x_i \le 0$, then this PSO is in IM(L). Otherwise, it is in IM(R).

Definition 8. PSO_k is prior to PSO_j if for PSO_k = <H(L), (x_k, y_k), (x_{ka1}, x_{ka2}), (y_{kb1}, y_{kb2})> and PSO_j =

<H(L), (x_j, y_j), (x_{ja1}, x_{ja2}), (y_{jb1}, y_{jb2})>, they satisfy one of the three prior conditions:

(a) $x_k < x_j$; (b) $x_k = x_j$ and $y_k > y_j$; (c) $x_{ka1} \le x_{ja1}$ and $x_{ka2} \ge x_{ja2}$; (d) $y_{kb1} \le y_{jb1}$and $y_{kb2} \ge y_{jb2}$;

We denote it as $PSO_k >> PSO_j$.

According to this priority, we describe PSO pattern (PSOP) of each image as the following form:

$$PSOP(IM_i) = <L_{i1}, L_{i2}, \ldots, L_{im}; R_{i1}, R_{i2}, \ldots, R_{in}>$$

Where L_{im} and R_{in} represent a PSO in $IM_i(L)$ and $IM_i(R)$ respectively.

Definition 9. PSOP of two images is complete similar if $|PSOP(IM_i)| = |PSOP(IM_j)|$, and the corresponding L_{ik} and L_{jk} (or R_{ik} and R_{jk}) have the same bound GS, locate the same part of the image (IM(L) or IM(R)) and satisfy the same prior condition. That is, if L_{ik} and $L_{i(k+1)}$ satisfy the prior condition (a) or (b), then L_{jk} and $L_{j(k+1)}$ must satisfy (a) or (b).

Definition 10. PSOP of two images is incomplete similar if cut one or more discontinuous PSO from one or both of these two images, then $|PSOP(IM_i)| = |PSOP(IM_j)|$, and the corresponding L_{ik} and L_{jk} (or R_{ik} and R_{jk}) have the same bound GS, locate the same part of the image (IM(L) or IM(R)) and satisfy the same prior condition.

For each image of one object, it has a PSOP and maybe is different from the PSOP of the adjacent image. We compare the PSOP of all adjacent images and get a pattern sequence in which two adjacent patterns are not same. This pattern sequence is called image sequence pattern (ISP).

The algorithm of discovering ISP is as follows.

```
DISP Algorithm:
Input: m images of one object
Output: Image sequence patterns (ISP)
1. Initialization: j=1, nj=1, k=1;
2. For i = 1 to m {
3. Compare the pattern PSOP(IMi) and
PSOP(IMi+1)of ith and (i+1)th image;
4. If complete similar
5. Then record the pattern as <
PSOP(IMk), nj = nj +1>;
6. Else if j = i
7.     Then k=i and record the pattern
as < PSOP(IMk), nj = 1>; j=j+1;
8.     Else k=i+1 and record the
pattern as < PSOP(IMk), nj = 1>;
j=j+1;}
```

We use an example to illustrate ISP from DISP algorithm, see figure 4. Convenient for description, we denote ISP as the following form:

$$ISP(S_i) = \{<M_i(IM_{i1}), n_1>, <M_i(IM_{i2}), n_2>, \ldots, <M_i(IM_{ik}), n_k>\}, \text{ or}$$
$$\{<M_i, n_1>, <M_i, n_2>, \ldots, <M_i, n_k>\}$$

Where for j=1…k, M_j and M_{j+1} are the distinct PSOP. IM_{ij} is the first image with the PSOP M_j and n_j is the number of the adjacent images with the same PSOP M_j.

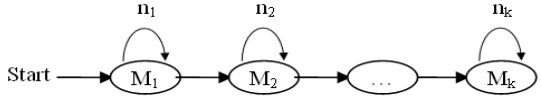


Figure 4: Patterns from Discovering ISP Algorithm

5.2 Mining ISSP of two objects

For $ISP(S_i)$ and $ISP(S_j)$, ISSP of two objects refers to the longest similar and continuous sub-patterns that belongs to the ISP of each object. For example, $ISP(S_i)$= {< M_1, 1>, < M_2, 2>, < M_3, 2>, < M_4, 1>, < M_5, 1>} and $ISP(S_j)$= {< M_1', 1>, < M_2', 3>, < M_3', 3>, < M_4', 2>, < M_5', 1>, < M_6', 1>}, if M_1 is similar to M_1', M_3 similar to M_3', M_4 similar to M_4' and M_5 similar to M_6', then ISSP of these two objects is < M_1', M_3', M_4', M_6' >. Here, M_3' and M_4' are called continuous sup-pattern, and M_1' and M_3' are discontinuous sup-pattern.

Since there are spatial relationship between all images in S_i, that is, for any IM_{i1}, IM_{i2}, …, IM_{in}, $IM_{i(j+1)}$ must be farther from calvaria than IM_{ij}, it is not necessary to retrieve the whole PSOPs of one object to find the similar pattern of a given M_i. For example, M_1 is PSOP of the farthest image from calvaria of one object and M_p is PSOP of the nearest image of another different object, it is not meaningful to compare M_1 and M_p to find whether they are similar or not because M_1 and M_p show the different parts of the brain. According to this, two rules are introduced to reduce the retrieval space and enhance the efficiency. For two objects S_i and S_j, assumed that $ISP(S_i) = m$ and $ISP(S_j) = n$,

(1) if m=n, that is, the number of the PSOPs in these two objects is equal, then we only need to retrieve M_{i-1}', M_i' and M_{i+1}' in $ISP(S_i)$ to discover the similar patterns of M_i in $ISP(S_i)$.

(2) If m<n, then we only need to retrieve M_i', M_{i+1}', …, M_{i+n-m}' in $ISP(S_i)$ to discover the similar patterns of M_i in $ISP(S_i)$.

The algorithm of discovering ISSP is as follows.

```
DISSP Algorithm:
Input: ISP of two objects
Output: Image sequence similarity
patterns (ISSP)
1. Initialization: C=NULL, NC=NULL;
2. Assumed that ISP(Si)=m and ISP(Sj)=n,
if m≤n, start the following steps:
3. For i = 1 to m {
4. if m=n, then goto step (5); if m<n,
then goto step(7);
5. Compare Mi in ISP(Si) with Mi-1', Mi'
and Mi+1' in ISP(Sj). If discover a
complete similar PSOP, C=C∪ (Mi→Mj). If
there is no PSOP to compare in ISP(Sj),
return step (3) to start the next
iteration. If there is no complete
similar PSOP to be discovered, goto step
(6);
6. Compare Mi in ISP(Si) with Mi-1', Mi'
and Mi+1' in ISP(Sj). If discover a
incomplete similar PSOP, NC=NC∪ (Mi→Mj).
If there is no PSOP to compare in
ISP(Sj), return step (3) to start the
next iteration;
7. Compare Mi in ISP(Si) with Mi',
Mi+1', …, Mi+n-m' in ISP(Sj). If discover a
complete similar PSOP, C=C∪ (Mi→Mj). If
there is no PSOP to compare in ISP(Sj),
return step (3) to start the next
iteration. If there is no complete
similar PSOP to be discovered, goto step
(8);
8. Compare Mi in ISP(Si) with Mi',
Mi+1', …, Mi+n-m' in ISP(Sj). If discover a
```

incomplete similar PSOP, NC=NC∪ ($M_i \rightarrow M_j$). If there is no PSOP to compare in ISP(S_j), return step (3) to start the next iteration; }
9. Order the whole patterns in C∪NC by the spatial relationship of M_i and use the exhausting method to discover the ISSP;

Since the above rules reduce the number of the PSOPs to be retrieved efficiently, the time cost of this algorithm mainly lies on the procedure of comparing the similarity of two PSOPs that depends on the count of the PSOs generated from the two algorithms in section 4.2. Therefore, the time complexity of this algorithm is O(km), where k is the number of the PSOs.

Definition 11. Two objects are similar if there exists ISSP in two objects.

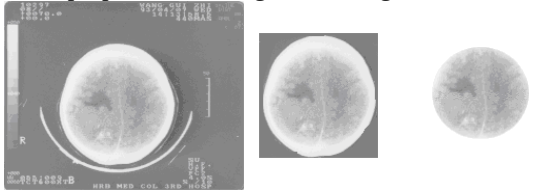
We give an example to illustrate this algorithm. For two objects S_i and S_j, ISP(S_i)= {< M_1, 1>, < M_2, 2>, < M_3, 2>, < M_4, 1>, < M_5, 1>, < M_6, 1>} and ISP(S_j)= {< M_1', 1>, < M_2', 3>, < M_3', 1>}, we use DISSP algorithm to get the results: C=($M_2' \rightarrow M_2$, M_4) and NC=($M_1' \rightarrow M_1$, M_3) ∪ ($M_3' \rightarrow M_6$), where C and NC includes the complete and incomplete similar patterns respectively. Notice that |ISP(S_j)|<|ISP(S_i)|, we order C∪NC by the spatial relationship of M_j in ISP(S_j). The result is ($M_1' \rightarrow M_1$, M_3) ∪ ($M_2' \rightarrow M_2$, M_4) ∪ ($M_3' \rightarrow M_6$). By means of exhausting method, the final ISSP is <M_1, M_2, M_6> and <M_3, M_4, M_6>. Therefore, S_i is similar to S_j.

6 EXPERIMENTS

To have access to real medical images is a very difficult undertaking due to legally privacy issues and management of hospital. But with some specialists' help and support, we got 103 pieces of precious data, which included 11 normal objects' CT scans and 92 abnormal objects data including CT scans and clinical data. Up to now, we have not found the method used on this kind of object dataset where each object has a brain image sequence. Therefore, we only give our experimental results.

6.1 Preprocessing with the guidance of DK1

For the preprocessing without the guidance of domain knowledge, the general method is cropping the images horizontally and vertically, see figure 5 (b). Figure 5 (c) is the preprocessed image with the guidance of DK1.



(a) Original Image (b) Cropped Image (c) Cropped Image with DK1

Figure 5: Digital brain image

6.2 Similarity Retrieval with the guidance of DK1

We randomly sample 10 percent of normal objects and 10 percent of abnormal objects as the targets from the dataset and use our algorithm to retrieve the similar objects of each target from the whole dataset. The formulas for precision and recall are given below to evaluate our algorithm:

$$P(p)=\frac{TP}{TP+FP} \qquad R(p)=\frac{TP}{TP+FN}$$

$$P(n)=\frac{TN}{TN+FN} \qquad R(n)=\frac{TN}{TN+FP}$$

Where TP stands for true positives, FP for false positives, FN for false negatives, TN for true negatives, P(p) and R(p) for the precision and recall of similarity retrieval of abnormal objects, P(n) and R(n) for the precision and recall of similarity retrieval of normal objects. Figure 6 and figure 7 shows the experimental results of randomly sampling 5 times. We can observe that the precision and recall are very high for normal targets. Only one abnormal object is retrieved as the similar object of normal targets for each time because the tumor in the brain of this abnormal object is laterally symmetrical. While our clustering algorithms calculate the difference set, they ignore the tumor. One normal object is not retrieved for each time because his (her) brain images are too bright and some noisy dark pixels cause the false results.

The average precision of the abnormal targets is more than 60% but not very high. All retrieved objects are certainly the abnormal objects with tumor in their brain and the image sequence of them is similar to that of the target. The reason why not very high is that their concrete tumor kind is not the same as the target and the medical doctors made a different detailed diagnosis. But this demonstrates that our similarity retrieval algorithm based on ISSP can gain the similar image sequence of the target. The average recall, however, is very high. This illustrates that the results of our similarity retrieval method based on ISSP can include most of the similar objects with the targets. For example, there are 7 objects including the target itself that are actually similar to the target in the dataset and our method can find 10 similar objects in which 6 objects belong to the actually similar objects.

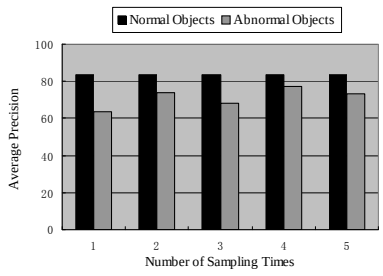


Figure 6: Average Precision of Similarity Retrieval when the Norman and Abnormal Objects as Targets

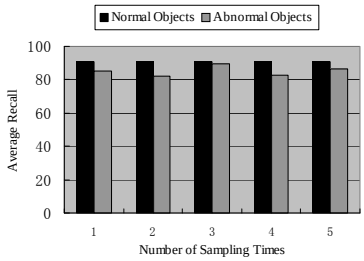


Figure 7: Average Recall of Similarity Retrieval when the Norman and Abnormal Objects as Targets

7 CONCLUSIONS

The high incidence of brain disease, especially brain tumor, has increased significantly in recent years. It is becoming more and more concernful to discover knowledge through mining medical brain image to aid doctors' diagnosis. In this paper, we firstly use two clustering algorithms to generate the possible Space-Occupying Lesion, and then discover ISP for each object. Next, we introduce a notion of image sequence similarity patterns (ISSP) for medical image database. ISSP refer to the longest similar and continuous sub-patterns hidden in two objects each of which contains an image sequence. These patterns are significant in medical images because the similarity for most medical images is not important, but rather, it is the similarity of objects each of which has an image sequence that is meaningful. We design a new algorithm with the guidance of the domain knowledge to generate ISSP for similarity retrieval. Our experiments demonstrate that the guidance of domain knowledge is meaningful and the results of similarity retrieval are interesting to medical doctors.

Future research includes further contact with medical knowledge that will make us engage in studying the classification, clustering and retrieval methods in medical images. Also, we will combine the clinical data with images to study new methods to enhance the accuracy of similarity retrieval and classification.Hope you find the information in this template useful in the preparation of your submission.

REFERENCES

[1] Zaiane, O.R. et al. (1998). Mining MultiMedia Data. CASCON: Meeting of Minds.

[2] WYNNE HSU, MONG LI LEE, JI ZHANG. Image Mining: Trends and Developments. Journal of Intelligent Information Systems, 19:1, 7–23, 2002.

[3] Vasileios Megalooikonomou, Christos Davatzikos, Edward H. Herskovits. Mining Lesion-Deficit Associations in a Brain Image Database. KDD-99 San Diego CA USA.

[4] Wynne Hsu, Mong Li Lee, Kheng Guan Goh. Image Mining in IRIS: Integrated Retinal Information System. Proceedings of the ACM SIGMOD, May 2000, Dellas, Texas, U.S.A., pp. 593.

[5] Y. Liu, F. Dellaert, W.E. Rothfus, A. Moore, J. Schneider, and T. Kanade. Classification-Driven Pathological Neuroimage Retrieval Using Statistical Asymmetry Measures. Proceedings of the Medical Imaging Computing and Computer Assisted Intervention Conference (MICCAI 2001), Utrecht, The Netherlands, October, 2001.

[6] Maria-Luiza Antonie, Osmar R. Zaiane, Alexandru Coman. Application of Data Mining Techniques for Medical Image Classification. Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001).

[7] Osmar R. Zaiane, Maria-Luiza Antonie, Alexandru Coman. Mammography Classification by an Association Rule-based Classifier. Proceedings of the Third International Workshop on Multimedia Data Mining (MDM/KDD'2002).

[8] Fayyad, U.M., Djorgovski, S.G., and Weir, N. (1996). Automating the Analysis and Cataloging of Sky Surveys. Advances in Knowledge Discovery and Data Mining, 471–493.

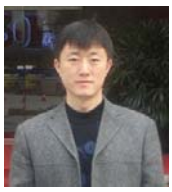
[9] Kitamoto, A. (2001). Data Mining for Typhoon Image Collection. In Second International Workshop on Multimedia Data Mining (MDM/KDD'2001).

[10] Ordonez, C. and Omiecinski, E. (1999). Discovering Association Rules Based on Image Content. In IEEE Advances in Digital Libraries Conference.

[11] Burl, MC et al. Mining For Image Content. In systems, Cybernetics, and Informatics / Information Systems: Analysis and Synthesis (1999).

[12] Soltanian-Zadeh H., Nezafat R., and Windham J.P.: "Is There Texture Information in Standard Brain MRI ?", Proceedings of SPIE Medical Imaging 1999: Image Processing conference, San Diego, CA, Feb.1999.

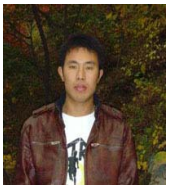
[13] Barra V., Boire J-Y., "Tissue segmentation on MR Images of the Brain by Positivistic Clustering on a 3D Wavelet Representation.", J. of Magnetic resonance Imaging, vol.11, pp. 267-278, 2000.



Haiwei Pan. I was born in July, 1974. I received my Ph.D. degree from the Department of Computer Science and Technology at Harbin Institute of Technology in 2006. My research interests include Parallel database, multimedia data mining, data warehouse and massive data process.

I am currently an Assistant Professor in the College of Computer Science and Technology at Harbin Engineering University.

I teach "Algorithm Design and Analysis" and "Graph Theory" for undergraduate students, and "Combinatorial Mathematics" for graduate students. I have published more than 20 Conference papers and Journal papers. My research has been supported by the National Natural Science Foundation of China, the Natural Science Foundation of Heilongjiang Province, the Fundamental Research Funds for the Central Universities.



Xiaolei Tan, born in 1982, M.S.candidate. His current research interests focus on data mining.

Qilong Han, born in 1974, Ph.d., associate professor. His current research interests include spatiotemporal database, protecting sensitive data, massive data process, real-time database.

Copyright © 2011 MECS