Original Article

CBMIR: Content-based Image Retrieval Algorithm for Medical Image Databases

Abdol Hamid Pilevar

Department of Computer Engineering, Bu Ali Sina University, Hamedan, Iran

ABSTRACT

We propose a novel algorithm for the retrieval of images from medical image databases by content. The aim of this article is to present a content-based retrieval algorithm that is robust to scaling, with translation of objects within an image. For the best result and efficient representation and retrieval of medical images, attention is focused on the methodology, and the content of medical images is represented by the regions and relationships between such objects or regions of the Image Attributes (IA) of the objects. The CBMIR employs a new model in which each image is first decomposed into regions. The similarity measurement between images is developed based on a scheme that integrates the properties of all the regions in the images using regional matching. The method can answer queries by example. The efficiency and performance of the presented method has been evaluated using a dataset of about 5,000 simulated, but realistic computed tomography and magnetic resonance images, from which the original images are selected from three large medical image databases. The results of our experiments show more than a 93 percent success rate, which is satisfactory.

Key words: *Content-based image retrieval, medical image databases, region matching*

INTRODUCTION

Content-based image retrieval (CBIR) applies to techniques for retrieving similar images from image databases, based on automated feature extraction methods. In recent years, the medical imaging field has been grown and is generating a lot more interest in methods and tools, to control the analysis of medical images. To support clinical decision-making, many imaging modalities, such as magnetic resonance imaging (MRI), X-ray computed tomography (CT), digital radiography, and ultrasound, are currently available. For administrative, clinical, teaching, and research activities, medical image database systems are emerging as an important component of Picture Archiving and Communication Systems (PACS). Usually, in the CBIR system, for each image, a feature signature on its pixel values is computed, the signature serves as an image representation, the components of the signature are called features. A rule for comparing images is defined as retrieving images that match the given query rules from a large database of images. The main reason for using the signature is to improve the correlation between image representation and semantics. This is done by mapping one or several signatures to d-dimensional points in some metric space and building an index on all signatures for fast retrieval. A function such as the Euclidean distance is used for calculating distances between each pair of signatures. The index is used to efficiently locate signatures close to the query point. The matched images are returned to the user.

The existing general-purpose CBIR systems roughly fall into two categories depending on the approach to extract signatures: The image-based search and the region-based search. Some of the systems using the weighted sum matching metric, combine the retrieval results from individual algorithms[1] or other algorithms.[2] The signatures are extracted; a comparison rule, including a querying scheme and the definition of a similarity measure between images is determined.

In most of the image retrieval systems, a query is specified by an image to be matched. We refer to this as an overall search, as similarity is based on the overall properties of images. By contrast, there are also partial search querying systems that retrieve based on a particular region in an image.[3] A content-based image retrieval method named CBDIR, segmenting the teeth of dental study models (plaster casts of the dentition), exhibits varieties of malocclusions.[4] A medical and general purpose image retrieval (MGIR) method is used for retrieving medical and general purpose images from databases, robust to scaling and translation of objects within an image.[5] The Colorimetry-Based Retardation Measurement Method (CBRM) is a method in which each

Address for correspondence:
Dr. Abdol Hamid Pilevar, Department of Computer Engineering, Medical Intelligence and Language Engineering Laboratory, Bu Ali Sina University, Hamedan, Iran. E-mail: pilevar@basu.ac.ir

image is first decomposed into regions. A measure for the overall similarity between images is developed using a region-matching scheme that integrates the properties of all the regions in the images.[6]

In a clinical decision-making system, more than a query by series ID, patient name, or study ID for images is needed. It is important and beneficial to find other images of the same disease and the same modality in the same anatomic region.[7-9]

In a query, for example, methods focusing only on color, texture, and shape, do not show how to handle inter-relationships or multiple objects or regions. The Content Based Medical Image Retrieval algorithm (CBMIR) algorithm mainly focuses on spatial relationships. The main contributions of this study are as follows:

1. A method for efficient retrieval and representation of medical images based on Image Attributes (IA)
2. An effective method for examining the retrieval process in the MRI and CT medical images.

The CBMIR system is interactive and the user is allowed to correct the results of the segmented images. The user can identify and extract interesting images or regions from all segmented images. The user can even specify the class to which an image belongs. Based on the properties of individual regions and spatial relationships between such regions, the CBMIR system takes the responsibility of efficient storage, representation, and retrieval of images.

The rest of this article is organized as follows:

- A short presentation of the underlying theory on Image Attributes is presented in Section 2
- An approach to the CBMIR algorithm for medical images is discussed in Section 3
- The indexing and search method is explained in Section 4
- Feature vector and similarity measure are discussed in Section 5
- Experimental results are discussed in Section 6
- The conclusion and issues for future research are presented in Section 7.

ATTRIBUTE EXTRACTION METHOD

A collection of images is given; the appropriate representations of their features and organization, together with their representations in the database are needed, so that one can search for images similar to the query image. The images are defined with their object properties and relationships between objects. The segmentation of CT and MRI images is in general very difficult and it is currently the subject of independent research activities.[1,10] The image segmentation process is done under the supervision of related experts (i.e., a clinician). In the first step, the images are segmented after necessary edge detection, using a low-pass filter. By editing, deleting or correcting the insignificant segments, the experts provide the desired segmentations and shapes.

Different features are specified for image representation; the features and original gray-level images are stored in the database and used for browsing or retrieving the images.

The images are segmented into disjointed regions or objects. Figure 1 shows an example of the edge-detected form of a gray-level image, and the complete contour-detected image is shown in Figure 1a, and its corresponding segmented polygonal shaped components are depicted in Figure 1 (P1 to P4).

The images are classified into specific predefined anatomical pathology classes [Table 1]. The classes are defined based on parts of the body (e.g., neck, head, etc.), or a part of the image (i.e., a region, an object, or a segment). They are classified into predefined classes, which correspond to the normality or abnormality of anatomical structures (e.g., hematoma, ventricle, tumor, etc.). The category is organized based on diagnostic and anatomical hierarchies. The experts have classified the images into appropriate anatomical classes by selecting their names in the class hierarchy. In this article, we focus on MRI and CT images, and the effectiveness of the proposed method is illustrated by looking at the image content, its representation, and retrieval. Figure 2 illustrates four representative images of these classes. However, the proposed method can be applied to other kinds of medical images too.

CBMIR ALGORITHM

Based on closed contour correspondence, the images are segmented into dominant image objects or regions, and then the image components are labeled by domain experts [Figure 3]. The classes of images are characterized by certain objects (e.g., liver, body outline, spine for CT or MRI images of the skull, abdomen, ventricles for images of the head etc.). For almost all of the images of the same class such objects are presented. The unexpected and additional objects are identified and classified into classes such as tumor, hematoma, and the like.

Three databases are implemented in this study:

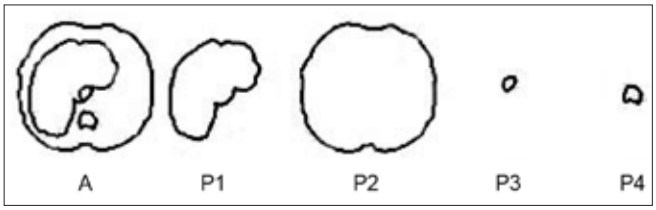

Figure 1: (A) Contour detection of an image. (P1, P2, P3, and P4) its segmented components

Table 1: Category of the anatomical pathology

A
- Aberrant subclavian artery
- Acalculia
- Acute biphenotypic leukaemia
- Adenoma
- Alcoholic liver disease
- Ann Arbor staging
- Arcuate uterus
- Atrophy
- Atypical teratoid rhabdoid tumor
- Auxesis (biology)

B
- Bone decalcification
- Bone healing
- Bone pathology
- Brain death
- Brain stem death
- British Neuropathological Society

C
- Cancer staging
- Carcinoma
- Cerebral arteriovenous malformation
- Cholesterolosis of gallbladder
- Circulating tumor cell
- CIT Program Tumor Identity Cards
- Cluster of differentiation
- Concentric hypertrophy
- Connective tissue neoplasm
- Cystic, mucinous, and serous neoplasms

D
- Dermatopathology

E
- Endocrine pathology
- Eosinophilic

F
- Flail limb
- Fluorescence *in situ* hybridization
- Frontotemporal lobar degeneration
- Frozen section procedure

G
- Gastrointestinal pathology
- Granulation tissue
- Granuloma
- Grinker myelinopathy
- Gross examination
- Gynecologic pathology

H
- Hamartoma
- Head and neck pathology
- High-intensity focused ultrasound
- Histopathology
- Hyperplasia
- Hypersegmented neutrophil
- Hypertrophy
- Hypoplasia

I
- Immunofluorescence
- Immunohistochemistry

L
- Leiomyoma
- Lesion

M
- MART-1
- Mature teratoma
- Morton's neuroma
- Myelolipoma

N
- Neoplasm
- Neuroectodermal tumor
- Neuropathology
- NK2 homeobox 1
- Nodule (medicine)

O
- Odontogenic tumor
- Ophthalmic pathology
- Osteoblastoma
- Osteoid osteoma

P
- Pediatric pathology
- Peliosis hepatis
- Perls' Prussian blue
- Pseudopolyps
- Pulmonary pathology

R
- Religious views on organ donation
- Renal pathology
- Rhabdomyosarcoma
- Romanowsky stain

S
- S-100 protein
- Sarcoma
- Schaumann body
- Soft tissue pathology
- Squamous-cell carcinoma
- Surgical pathology

T
- Teratoma
- Tuberculous lymphadenitis
- Tumor

U
- Unicornuate uterus
- Uterine artery embolization
- Uterine septum
- Uterus didelphys

V
- Vein of Galen aneurysmal malformations
- Visceroptosis

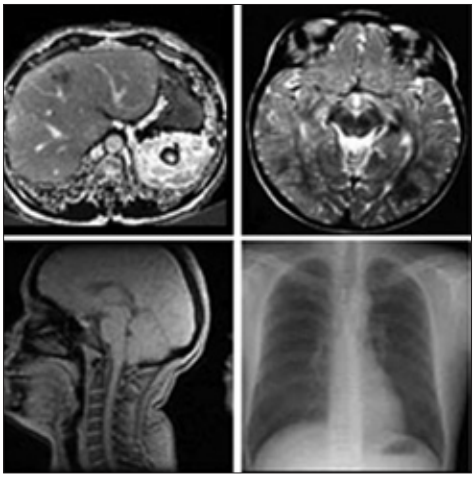

Figure 2: Four samples of the medical images which are saved in training and test databases

- All the original MRI or CT images saved in a database are named Dd
- Images of the closed contour components (i.e., Figure 1 P1…P4), saved in a database are named Ds
- ID numbers of the images and numbers of their components are saved in database Dy.

This is followed by the number of components and feature vectors of the images, which are saved in database Dv.

Step 1: CBMIRs algorithm for creating the Dv database:

for k=1: N */ N is the number of images in database Dd

read image k; */read the kth image from database Dd

for I=1: P_k */ P_j is the number of components of the k^{th} image

read i; */ read component i, from database Ds (i.e., Figure 3 P_i)

calculate C_i; */ C_i is the centroid point of the component region i

find S_i; */ S_i is the signature of component i (i.e. Figure 3 S_i)

calculate F_{i1}, F_{i2}… F_{i8}

*/ calculate eight feature values for F_i, of signature S_i, */ on points (0, $\pi/4$, $\pi/2$, $3\pi/4$, π, $5\pi/4$, $3\pi/2$, $7\pi/4$) (i.e.

Figure 3 S_i)
save k, i, F_{i1}, F_{i2} …F_{i8}
*/ save ten values for component i, in database D_v
end;
end;

Step 2: CBIR-Ms algorithm for query matching:
input Q; */Q is the query image, segmented into closed contours form
find F_q;*/ calculate the feature vector of query image Q, like step 1
for J=1: N */ N is the number of images in database D_d
compare Q; */ search for the most similar images in database D_y
*/ (compare number of components and feature vectors)
report similarities; */ report partially or completely similar cases.
end;

All the images in the database are normalized into 400 × 400 pixel size images, to reduce the fuzziness of the attributes.

Two images are compared by the system for object similarities and their relationships, and outcome results are evaluated by referring to instructions of the radiologists.

INDEXING AND SEARCH METHOD

The images are coded with character strings of a length of six characters, as shown in Figure 4. For locating an image address in database Dd, these six character string codes, which are called class-codes, are used.

To search for the feature vectors of the components in databases Dy, string codes with the length of eight characters are used [Figure 5].

A special tree search technique called component feature codes (CF-Codes) is applied; the feature vectors are looked for by using a tree-based searching technique as displayed in Figure 6.

For example, a query image is given and its CF-code is extracted by the system, as shown in Figure 7, then the searching process will be followed through the nodes of the tree, as marked in Figure 6.

An appropriate software system is provided, every time a query image is given to the system, and the following steps are taken:

1. Image is segmented into components
2. The number of components is found
3. Contours of the components are detected
4. Signatures of the components are found
5. The feature vector of the components is computed
6. The class, sub-class, and sub-sub-class are suggested and class-code is extracted from database D_d

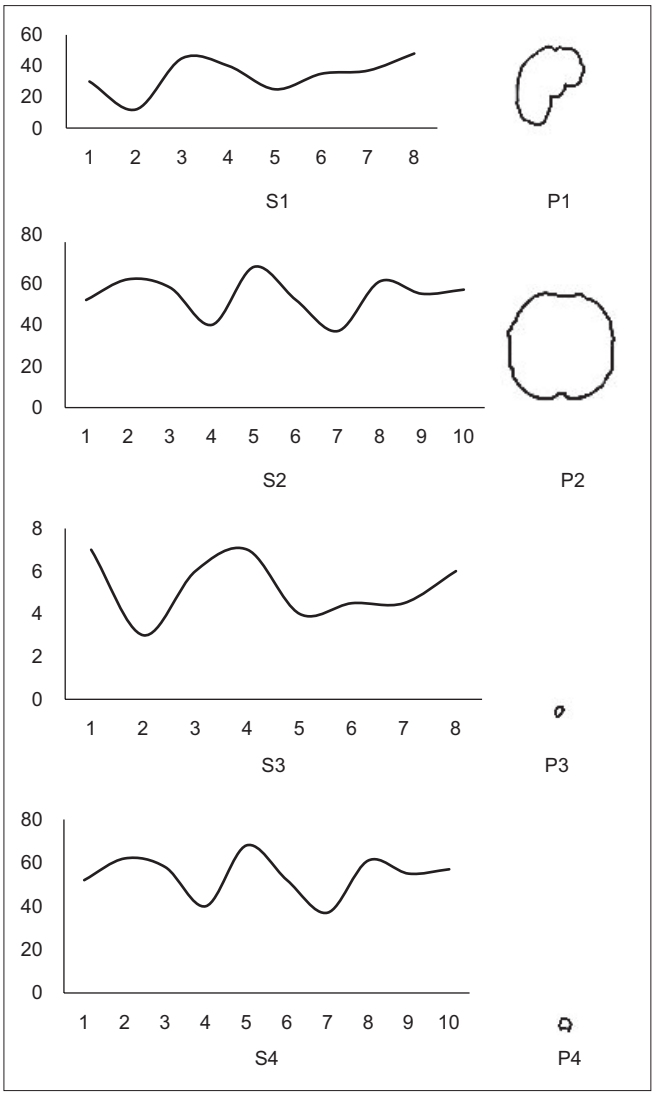

Figure 3: A sample of extracted components P1…P4, and their signature S1…S4

Class: 2 characters	Sub-class: 2 characters	Sub-sub-class: 2 characters

Figure 4: The structure of character string codes (class-code) of the images in database D_d

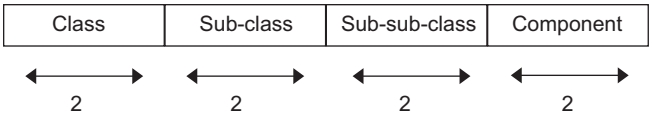

Figure 5: The structure of character string codes of the features in database D_y

7. The CF-code of the image is provided
8. The records with similar CF-codes in database D_y are marked
9. The records with the most similar feature vectors in database D_y are found
10. The related images in database D_d that compare to

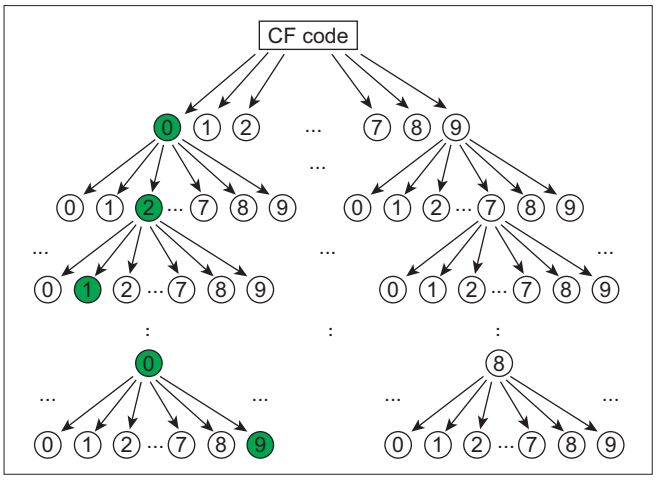

Figure 6: Depict of CF-codes tree base search technique

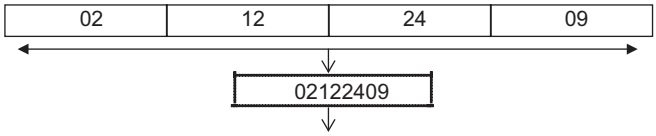

Figure 7: An example of a CF code, consisting of class-code (02), sub-class-code (12), sub-sub-class-code (24), and component-code (09)

marked records in database D_y are detected

11. The detected images are reported respectively, based on their similarities.

FEATURE VECTOR AND SIMILARITY MEASURE

Features are extracted as follows:
Let I be the Image, V_f be vector defined for feature *f*, then function between I and V_f is defined as (1).

$$E_f: I \rightarrow V_f \tag{1}$$

For any extracted feature there exists a feature vector into which the images are mapped. All images are known by their related feature vector. Therefore, too many feature vectors have to be supported by the image database system, while different measurement strategies can be applied. There are different metric functions for determining the similarity degree of the images with each other. A metric vector is defined as a tuple (V_f, M_f), where V_f is a set of features and M_f a metric for calculating the similarity between a pair of given features V_f, as follows:

$$M_f: V_f \times V_f \rightarrow \mathrm{R} \tag{2}$$

Where $V_f \times V_f$ is the Cartesian product between the features of the same vector.

Such that:

1. $M_f(x, y) \geq 0$. Non-negativity
2. $M_f(x, y) = 0$, if and only if x = y. Identity
3. $M_f(x, y) = M_f(y, x)$. Symmetry
4. $M_f(x, z) \leq M_f(x, y) + M_f(y, z)$. Triangle inequality

If a linear combination of different metrics and many features are used, then better comparisons are expected to be achieved:
- Let E_f be the feature extraction function of a feature *f*.
- Let x,y ∈ I be images.
- Let M_f be a metric in the feature space V_f.

If the importance weight is defined as W_f, and linear combination of metrics as M_f, then the similarity function of features is defined as:

$$d(x, y) = \sum_f w_f M_f(E_f(x), E_f(y)) \tag{3}$$

Whereas in this study, the number of features in each feature vector is eight, therefore:

$$\sum_{k=1}^{8} w_k = 1$$

Measures can be evaluated using distance functions, and it is important to determine the most suitable function for each type of feature vector. The following metrics have been experimentally evaluated. Let *H, H′* be vectors, both with M elements, and H_m be the m^{th} individual element.

Euclidean distance is used to evaluate distances in n dimensional vector spaces,

$$L(H, H') = \left(\sum_{m=1}^{M} (H_m - H'_m)^2 \right)^{\frac{1}{2}} \tag{4}$$

In this study, the number of the elements in vector *M* is set to be eight features.

Ten of the most similar images to the query images are selected from database D_d. Between thses ten selected images, the most similar image is found by using the Fourier descriptors method as follows:

The Fourier descriptors can be used to match similar shapes even if they have a different size and orientation. If a(f) and b(f) are the FDs of two boundaries u(n) and v(n), respectively, then their shapes are similar if the distance,

$$d(u_0, a, \theta_0, n_0) \triangleq \min_{u_0, a, \theta_0, n_0} \left\{ \sum_{n=0}^{N-1} | u(n) - av(n + n_0) e^{j\theta_0} - u_0 |^2 \right\} \tag{5}$$

is small. The parameters u_0, α, n_0, and θ are chosen to minimize the effects of translation, scaling, starting, points, and rotation, respectively. If u(n) and v(n) are normalized so that ∑u(n) = ∑v(n) = 0, then for a given shift n_0, the above distance is minimum when

$u_0 = 0$

$$a = \frac{\sum_k c(k)\cos(\psi_k + k\Phi + \theta_0)}{\sum_k |b(k)|^2} \quad (6)$$

And

$$\tan\theta_0 = \frac{\sum_k c(k)\sin(\psi_k + k\Phi)}{\sum_k c(k)\cos(\psi_k + k\Phi)} \quad (7)$$

Where: $a(k)b^*(k) = c(k)e^{j-k}$, $\Phi \triangleq -2\pi n_0/N$, and c(k) is a real quantity. These equations gives α and θ_0, from which the minimum distance d is given by

$$d = \min_\Phi[d(\Phi)] = \min_\Phi \left\{ \sum_k |a(k) - \alpha b(k)\exp[j(k\Phi + \theta_0]|^2 \right\} \quad (8)$$

The distance $d(\Phi)$ can be evaluated for each $\Phi = \Phi(n_0)$, $n_0 = 0, 1,\ldots,N-1$, and a minimum search to obtain d. The quantity d is then a useful measure of difference between the two shapes.

The image with minimum difference is reported as the most similar image to the query image.

EXPERIMENTS AND RESULTS

A dataset of about 5,000 simulated, but realistic computed tomography and magnetic resonance images (MRI) is used. In the medical field, there is a great amount of anatomical information gathered during the past centuries, which is well accepted by physicians and radiologists. The original images are selected from three different and large medical image databases, such that almost all classes of the category 'anatomical pathology' [Table 1] are covered.

As an example, a query image is given to the CBMIR's software system and the features are extracted as shown in Table 2.

The feature vector of the query image is compared with the feature vectors of the images in the dataset. The Euclidian method is used to calculate the distances between the query image and extracted images. The most similar images are reported. In Table 3, the features vector of the most similar extracted image is shown.

The two primary measures for evaluating the overall performance of a retrieval system are recall and precision. Recall is defined as the fraction of retrieved relevant images over the total number of relevant images in the database. Precision is defined as the fraction of relevant images retrieved over all the images retrieved by the system. The two measures are usually correlated in such a way that maximizing one deteriorates the other. For each query, the precision is computed and the accuracy of the method is measured.

Our experiments on the CBMIR system illustrate that, the average value of precision is more than 93%, that is to say, more than 93 percent of the retrieved images are correct.

For example in Figure 8, a query image with four components is shown [Figure 8a]. One of the extracted images by the CBMIR algorithm is image 2 with six components [Figure 8b], as we can see, four of the components are similar to the query image and two components are extra. It can be very noticeable for experts, from a diagnostic point of view.

CONCLUSION

In this article, we have proposed a novel content-based retrieval algorithm CBMIR, robust to translation and scaling of objects within an image. CBMIR employs a novel technique in which each image is first decomposed into components. An efficient software system is used for illustrating the signatures and computing feature vectors. CBMIR, unlike the traditional approaches, which are based on a single signature for each image, builds a set of eight signatures for an image and stores the set of signatures as feature vectors in a grid-structured file. Experiment results on real-life sets show that the retrieved images by CBMIR are semantically more related to the query

Table 2: Feature vectors of a query image with five components

Image ID/ Component	Feature							
	1	2	3	4	5	6	7	8
142109/1	12	5	7	18	15	3	7	5
142109/2	3	15	4	10	12	13	16	7
142109/3	7	11	12	9	6	8	10	20
142109/4	15	6	10	7	15	10	11	11
142109/5	10	7	9	3	9	7	5	15

Table 3: Feature vectors of the most similar images to the query image

Image ID/ Component	Feature							
	1	2	3	4	5	6	7	8
Query/1	9	5	7	17	15	3	7	8
Query/2	3	13	4	10	12	13	16	7
Query/3	7	8	12	9	6	8	10	18
Query/4	15	6	10	7	13	14	11	11
Query/5	15	7	9	4	12	7	5	12

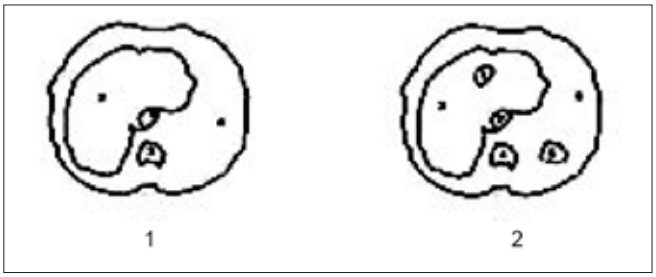

Figure 8: (1) Query image with four components, and (2) extracted image with six components

image than those retrieved by similar algorithms. In this article we have shown that the CBMIR system works efficiently and effectively for medical applications.

In future studies, we have planned to work on applying this technique to different types of large and very large medical and biomedical image databases, which will be presented in our next article.

REFERENCES

1. D. Comaniciou, P. Meer, and D. J. Foran. Image-guided decision support system for pathology. Mach Vision Appl 11, pp. 213-23, 1999.
2. S. C. Orphanoudakis, C. E. Chronaki, and D. Vamvaka. I-Cnet: Content-based similarity search in geographically distributed repositories of medical images. Comput Med Imaging Graph 20, pp. 193-207, 1996.
3. E. G. Petrakis, and C. Faloutsos. Similarity searching in medical image databases. IEEE Trans Knowl Data Eng 9, pp. 435-47, 1997.
4. A. H. Pilevar, M. Sukumar, A. R. Gowda, and E. T. Roy. CBDIR: Content-based dental image retrieval. J Indian Orthod Soc 6, pp. 38-52, 2005.
5. A. H. Pilevar, and M. T. Pilevar. MGIR: An image retrieval method for medical and general image databases. *11th Joint Conference on Information Science (JCIS 2008)*, Dec. 15-20, 2008.
6. A. H. Pilevar, and M. T. Pilevar. CBRM: A new content-based regional-matching image retrieval method. *Proceeding of the eighth IASTED international conference, visualization, imaging and image processing*, Sept.1-3, 2008.
7. K. P. Andriole. The Society for Computer Applications in Radiology Transforming the Radiological Interpretation Process (TRIP) Inititative. *White Paper (http://www.siimweb.org)*. November 2005. [Last accessed on March 26, 2007].
8. A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. IEEE Trans Pattern Rec Machine Intell 22, pp. 1349-80, 2000.
9. A. A. Youssif, A. A. Darwish, and R. A. Mohamed. Content based medical image retrieval based on pyramid structure wavelet. Int J Comput Sci Netw Secur 10, pp. 157-64, 2010.
10. W. W. Chu, C.-C. Hsu, A. Cardenas, and R.K. Taira. Knowledge-based image retrieval with spatial and temporal constructs. IEEE Trans Knowl Data Eng 10, pp. 872-88, 1998.

How to cite this article: Pilevar AH. CBMIR: Content-based image retrieval algorithm for medical image databases. J Med Sign Sens 2011;1:12-8

Source of Support: Nil, **Conflict of Interest:** None declared

BIOGRAPHY

Abdol Hamid Pilevar, Assistant professor in Computers Engineering Department, Bu Ali Sina University, Hamedan, Iran. He received his B.Sc and MSc. degrees in Computer Systems from Florida Atlantic University, Florida, U.S.A.

Dr. Pilevar received the PhD degree in computer science from the University of Mysore, Mysore, India in 2005. He was a Research Associate and Post Doctoral Fellow at the Indian Institute of Science (IISC, Bangalore, India), and Mediscan Prenatal Diagnosis and Fetal Therapy Center (Chennai, India) in 2005 – 06.

His fields of interest include Medical Intelligence and Image Processing, 3D Modeling, and Speech and Natural Language Processing. He has published more than 40 articles in international and national journals, and conferences.