Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2013) 21: 882 – 896

doi:10.3906/elk-1010-876

A framework for medical image retrieval using merging-based classification with dependency probability-based relevance feedback

Hossein POURGHASSEM,[1,*] **Sabalan DANESHVAR**[2]

[1]Department of Electrical Engineering, Najafabad Branch, Islamic Azad University, Isfahan, Iran

[2]Faculty of Electrical Engineering, University of Tabriz, Tabriz, Iran

Received: 22.10.2010 • **Accepted:** 10.04.2011 • **Published Online:** 03.05.2013 • **Printed:** 27.05.2013

Abstract: Content-based image retrieval (CBIR) systems are used to retrieve relevant images from large-scale databases. In this paper, a framework for the image retrieval of a large-scale database of medical X-ray images is presented. This framework is designed based on query image classification into several prespecified homogeneous classes. Using a merging scheme and an iterative classification, the homogeneous classes are formed from overlapping classes in the database. For this purpose, the shape and texture features, selected using the forward selection algorithm, are optimized by a novel genetic algorithm-based feature reduction and optimization algorithm in the feature space. In this algorithm, using a new fitness function, we try to locate similar images in the database together in the feature space. Using the merging-based classification, the m-nearest classes to the query image are selected as a filtered search space. To increase the retrieval efficiency, we integrate a novel dependency probability-based relevance feedback (RF) approach with the proposed CBIR framework. The proposed RF uses a synthetic distance measure based on the weighted Euclidean distance measure and Gaussian mixture model-based dependency probability similarity measure of the database images to the Gaussian mixture distribution function of the positive images. The experimental results are reported based on a database consisting of 10,000 medical X-ray images of 57 classes (ImageCLEF 2005 database). The provided results show the effectiveness of the proposed framework compared to the approaches presented in the literature.

Key words: Content-based image retrieval, merging-based classification, dependency probability-based relevance feedback, medical X-ray images, genetic algorithm-based feature reduction and optimization algorithm

1. Introduction

Content-based image retrieval (CBIR) is an additive needed for search and retrieval in image databases. To index images for image retrieval applications, we use low-level features such as the colors, textures, and shapes of objects to represent the image content [1]. Even though CBIR systems have been applied widely in general applications (such as face and fingerprint matching for identification, Internet shopping, and logo searching), only a few CBIR systems (such as ASSERT [2], IRMA [3], and NHANES II [4]) have been extended specifically for medical applications. Many medical CBIR systems, such as high-resolution computed tomography (HRCT) lung images [5,6], mammography [7], chest computed tomography [8], chest X-ray [9], spine X-ray [4,10–12], and dental X-ray [13], often present images with specific organ and modality or diagnostic study that cannot be used in other medical applications. A few systems have been developed for general medical application (e.g., MedGIFT [14], KmED [15], and IRMA [3]). For example, in the Gaussian mixture model-Kullback–Leibler

*Correspondence: h_pourghasem@iaun.ac.ir

(GMM-KL) framework presented in [16], using a probabilistic image representation scheme based on the GMM and an image-matching strategy based on the KL measure, a classification-based image retrieval framework was designed. In this framework, a class model is extracted based on the GMM estimation of the images, and then it is used to categorize and retrieve the database images. In [17], using a probabilistic multiclass support vector machine (SVM) and fuzzy c-mean clustering for categorizing of images, a CBIR framework for medical images was introduced. In the presented relevance feedback (RF) of this framework, using the relevant and irrelevant images, the most dominant negative and positive categories are predicted by SVM and the voting rule. Next, the images of the most dominant positive class are rewarded and the images of the most dominant negative class are punished. In [18], a classification-based medical image retrieval framework was designed using category membership scores, a combination of probabilistic classifiers, and an adaptive similarity fusion scheme, and a feature-level fusion was applied. In [19], a classification-based medical image retrieval method for a special medical database was presented.

In this paper, a CBIR framework is presented for medical X-ray image applications. We have 2 principle problems in the medical X-ray image classification problem. First, there is an intense overlapping between different images of different classes in very large databases. Second, some classes of the database have an intense intraclass distance [1] based on body orientation, anatomic region, and texture contents in our CBIR framework. Therefore, this framework has been designed based on the merging-based classification of the query image into several prespecified homogeneous classes using shape and texture features. Using the proposed merging scheme in [1], the homogeneous classes are formed from overlapping classes in our database. For this purpose, the selected features are optimized and reduced by a novel genetic algorithm-based feature reduction and optimization (GFRO) algorithm, until more semantic classes of the similar images are formed. In other words, the GFRO algorithm modifies the feature space until more similar images are located together.

In this paper, a novel RF approach has been integrated into our framework for the improvement of the retrieval performance. The proposed RF approach is a novel dependency probability-based approach based on the combination of the weighted Euclidean distance and the GMM-based dependency probability similarity measures. In each iteration of the proposed RF, the weights of these 2 measures in the synthetic measure are determined based on their retrieval performance in the previous iteration.

The rest of this paper is organized as follows: the details of the proposed framework for content-based medical X-ray image retrieval are described in Section 2, where the proposed feature extraction, selection, and optimization stages and the merging-based classification are described. The details of the proposed dependency probability-based relevance feedback approach are presented in Section 3. Section 4 contains the experimental results of the proposed framework. A discussion and a conclusion on our proposed framework and the work are contained in Sections 5 and 6, respectively.

2. The proposed framework for content-based medical X-ray image retrieval

The block diagram of the proposed framework for content-based medical X-ray image retrieval is shown in Figure 1. In this block diagram, features are extracted from the query image, and then the m-nearest classes to the query image are determined by the merging-based classifier. Similar images in the search space are sorted by the similarity measure and are presented to the user. Positive and negative images in the presented images are labeled by the user, and are then used to improve the retrieval performance by a RF algorithm. In this section, the details of feature extraction and the merging-based classifier are described.

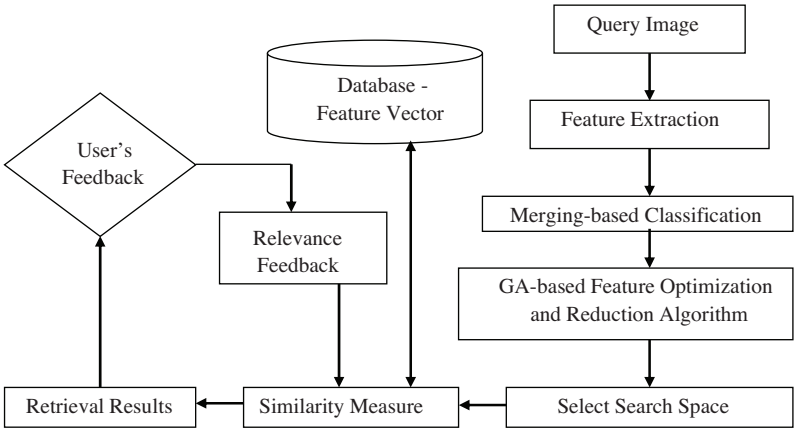

Figure 1. Block diagram of the proposed framework for content-based medical X-ray image retrieval.

2.1. Feature extraction

Image representation by low-level features has a major effect on the performance of a CBIR system. Many classes of our database consist of images with different objects whose shape features can be used to distinguish them. Therefore, we have extracted the first 5 invariant moments [20] of the main object in the binary image. To extract the main object, the X-ray image is binarized using Otsu's thresholding algorithm [21], so that the binary image is divided into a background and one or more objects as a foreground. If the foreground of the binary image contains more than one object, connect component analysis is used to label its objects based on pixel connectivity (e.g., 8-connected). The object that has the most area is extracted as the main object. Based on central moments with different orders, a set of moments invariant to translation, rotation, and scale can be derived. Moreover, we have extracted Fourier descriptors [22] as a shape feature that describe the shape of an object with the Fourier transform of its boundary. In this representation, the boundary pixels of the main object have been represented based on the complex coordinate. Next, Fourier descriptors are defined based on Fourier transform of the complex coordinate representation. We resample the boundary of each object to M samples with a uniform sampling function for equalizing the length of the extracted shape features of all of the objects in our database. In our application, the number of samples is set to 256.

The major axis orientation, eccentricity, and major and minor axis length features [23] have also been extracted from the binary object. The direction of the largest eigenvector of the second-order covariance matrix of an object is considered as the major axis orientation. The major and minor axes of the ellipse that have the same normalized second central moments with an object are noticed as the major and minor axis lengths, respectively. The ratio of the smallest eigenvalue to the largest eigenvalue is considered as eccentricity [23]. Therefore, the shape feature vector includes a total of 263 elements (invariant moments = 5, Fourier descriptors = 254, major axis orientation, eccentricity, and major and minor axis lengths = 4).

The texture features of medical images of different classes with similar shape content can be used to distinguish these classes. In this paper, we extract the texture features from the gray-level cooccurrence matrix [24]. A gray-level cooccurrence matrix is defined based on the occurrence probability of the different gray levels of 2 pixels that have a displacement d and an angle θ together. The contrast, homogeneity, energy, and correlation features are measured based on the gray-level cooccurrence matrix. The gray-level cooccurrence matrix for 4 different directions ($\theta \in \{0^\circ, 45^\circ, 90^\circ, \text{ and } 135^\circ\}$) and the distance $d = 1$ is calculated. Therefore, a 16-element vector constitutes the texture feature. In [1], a 2-directional histogram and tessellation-based

spectral features in wavelet transform subbands were proposed that were customized for medical X-ray image classification. In our application, we have also extracted them. The directional histogram feature captures spatial information from the subbands of the wavelet transform [1]. In this paper, 2-level Daubechies wavelet transforms are applied and the number of quantization levels is set to 3. Therefore, a 36-element vector is formed by the directional histogram feature. The tessellation-based spectral feature demonstrates the content of the image based on the edge and orientation information as well as the coarseness of the objects to apply a circular tessellation scheme on the frequency spectrum of the accumulated image of the 2 modified horizontal and vertical wavelet transform subbands [1]. In this paper, the feature vector of the tessellation-based spectral feature is formed based on the standard deviation within the sectors of the tessellation scheme. Therefore, this feature vector is a 16-element vector. Finally, a 331-element vector consisting of the shape and texture features is formed.

2.2. The merging-based classification

In this paper, a classification-based image retrieval algorithm is developed. Therefore, the classification performance of the medical image is very important in the performance of the image retrieval system. In our application, an ordinal classifier with constant features cannot classify images with high performance, due to the major overlapping of the many classes in our database. Therefore, to increase the classification performance and create homogeneous classes based on body orientation, anatomic region, and texture contents, we have 2 problems. First, to increase the interclass distance between different classes and also decrease the intraclass distance of each class, which features should be applied in the feature vector? Second, to improve the classification performance and also construct semantic classification based on the body orientation and anatomic region, which classes should be merged together? To solve these problems, first the images are classified by a multilayer perceptron (MLP) classifier, and then the forward selection algorithm is used to select the best feature vector [25]. Ultimately, the weight of each element of the selected feature vector is determined by the GFRO algorithm. Secondly, using the proposed merging scheme in [1], homogenous classes are formed based on merging the overlapping classes.

2.2.1. Feature selection

In the feature selection procedure, we try to find a minimum and optimal set of the extracted features that obtains the best classification performance. Since this optimal set of features is unknown, usually 2 common forward selection and backward elimination algorithms are utilized [25]. Due to the high computational complexity of the backward elimination algorithm in regard to the forward selection algorithm [26], we apply the forward selection algorithm as a feature selection algorithm.

2.2.2. The proposed GFRO algorithm

To improve the performance of our classification-based image retrieval algorithm, we propose a GFRO algorithm to determine the weight of each element in the feature vector. In this algorithm, a new fitness function is defined based on the number of similar images in the K-nearest neighbor of each image of the database images in the feature space. In other words, it reconstructs the feature space until more similar images are located together in the feature space. Using this algorithm, the weight of each feature in the feature vector is determined and, ultimately, each element of the feature vector with a weight that is smaller than the threshold value (T_{GA}) is removed from the feature vector as a weak feature. The optimum threshold of T_{GA} is determined by trial and

error. If T_{GA} is set to higher or lower values, then it may cause the elimination of strong features or the adding of weak features to the feature vector. Indeed, this algorithm not only optimizes the feature vector, but also reduces its length.

In our algorithm, the weights of the feature vector are determined based on a chromosome representation [27]. In other words, the length of the feature vector and the number of genes in the chromosome are equal. In our algorithm, the fitness function has a vital role because it selects individuals (the weights of the features) to regenerate the next generations. Using the individual's fitness, the better individuals with more of a chance will be selected based on a probabilistic measure. We define the fitness function based on the number of similar images in the K-nearest neighbor of each image of the database images in the feature space. In other words, the individual's fitness is improved if similar images are located together in the feature space. The probability, P, for each individual is defined by:

$$P = \frac{1}{T}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(\frac{n_{ij}}{K}), \tag{1}$$

where M and N_i are the number of classes in the database and the number of images in the ith class, respectively. $T$and n_{ij} are the total number of images in the database and the number of similar images in the K-nearest neighbor of the image j in the class i, respectively. Eq. (1) shows that the total number of database images is applied for calculating the fitness value of each individual. In our application, $K = 17$ is set by trial and error. The weighted Euclidean distance between image I and image D in the feature space is calculated by:

$$Dis(I,D) = \sqrt{\sum_{i=1}^{d} w_i(I_i - D_i)^2}, \tag{2}$$

where $d$is the length of the feature vector and w_i is the assigned weight to the feature vector ($(w_1, .., w_i, .., w_d)$ are elements of a chromosome). Note that the larger fitness value is determined by the better chromosome.

In the genetic algorithm, genetic operators such as arithmetic crossover, heuristic crossover, simple crossover, and nonuniform and uniform mutation provide new solutions in the next generation using existing solutions in the current population [27]. The arithmetic crossover operator creates new individuals that are the weighted arithmetic mean of 2 parents. The parent with the better fitness value has more weight in the weighted arithmetic mean calculation. The heuristic crossover operator creates children based on the fitness values of 2 parents on the line they are in. In other words, the new child is located a small distance (20% of the Euclidean distance between 2 parents) from the better parent. The simple crossover operator creates a new child to incorporate the 2 entries that these 2 entries have been selected from, i.e. 2 parents randomly. In our application, the elite count and crossover fraction parameters are set to 2 and 0.5, respectively. The nonuniform mutation is applied to a Gaussian distribution centered on 0, with scale and shrink parameters set to 1, while uniform mutation is used with a mutation rate 0.15.

The initial population is randomly generated in the range of [0, 1] with 30 chromosomes as the population size. The genetic algorithm will be stopped if there is no improvement in the fitness function for 40 consecutive generations, if the sum of the deviations among the individuals becomes smaller than 1e-6, or, ultimately, if the maximum number of iterations (i.e. 100 iterations) is carried out [27].

2.2.3. The merging scheme

Using the merging scheme presented in [1], we merge the classes together with maximum overlapping. The merging scheme introduced in [1] has an iterative mechanism for merging the overlapping classes, i.e. in each iteration of the merging scheme, the overlapping classes are merged together, and then using a MLP classifier, the decision boundaries in the feature space are reconstructed based on the newly merged classes and the total accuracy rate of the classification is calculated on the test dataset. The iterative procedure of the merging scheme will be continued until the desired total accuracy rate ($T_{desired}$) value is satisfied. The merging scheme applies 3 measures to recognize the overlapping classes: the accuracy rate, misclassified ratio, and dissimilarity. The accuracy rate and misclassified ratio measures are defined based on the classification results. However, the dissimilarity measure is defined based on the correlation distance of the distribution functions of 2 classes [1]. Class i and class j will be merged if they meet 3 conditions. First, the accuracy rate of class i is less than a predefined threshold (λ); second, the misclassified ratio of 2 classes is more than a predefined threshold (β); and third, the dissimilarity of 2 classes is less than a predefined threshold (γ). The selection of the values of the 3 thresholds, λ, β, and γ, has a critical role in the merging scheme performance. The optimal values of the 3 thresholds, λ, β, and γ, are determined based on the application and opinion of the user in regard to the desired number of classes and the acceptable total accuracy rate. However, if we want to increase the total accuracy rate and the number of classes that have the merging conditions, we should set the higher values to λ and γ and the lower values to β [1]. The optimal values of the 3 thresholds, λ, β, and γ, are determined based on forming the homogeneous classes.

3. The dependency probability-based RF

RF approaches are used to improve the performance of CBIR systems. Query-point moving and weight updating are 2 procedures that are used to construct most of the RF approaches. In the query-point moving approach, the ideal query point estimation is improved by moving the current query point [28]. In the weight updating approach, the weights used in the computation of the similarity measure are modified based on the user's feedback [29].

The proposed RF approach is a novel synthetic approach based on the combination of the weighted Euclidean distance measure and the GMM-based dependency probability similarity measure. The weighted Euclidean distance measure is defined based on the weighted Euclidean distance between the query image and database images, whereas the GMM-based dependency probability similarity measure is the dependency probability of the database images to the Gaussian mixture distribution function of the positive images (Figure 2). In this paper, as in [30], GMM is used to estimate the distribution function of the positive images. In [30], the proposed RF approach was executed based only on the dependency probability of the database images to the GMM of the positive images, while we combine this measure with the weighted Euclidean distance measure using a novel combination algorithm. In the block diagram of Figure 2, to reduce the computational complexity of the GMM-based distribution function estimation, feature vectors of the positive images have been reduced to 8 elements using the principle component analysis (PCA) algorithm [31]. Next, the distribution function of a class is defined based on the GMM of the positive images of such a class. The Gaussian mixture distribution function is defined as:

$$f(x|\theta) = \sum_{i=1}^{Z} \eta_i \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\}, \tag{3}$$

where x is a feature vector, the η_is represent the mixing weights ($\sum_{i=1}^{Z} \eta_i = 1$), θ includes the collection of parameters μ_i and Σ_i, and d represents the dimension of the feature space. Using the expectation-maximization algorithm, the maximum likelihood parameters of a mixture of Z Gaussians in the feature space is determined [16]. By evaluating the proposed RF performance, we determine the optimal number of Gaussian functions in the GMM. The GMM-based dependency probability similarity value of the extracted feature vector x of the query image to given class parameter θ is determined based on the value of $f(x|\theta)$. The maximum value of $f(x|\theta)$ indicates more dependency probability values of the query image x to the class parameters θ. The synthetic distance measure is defined as below:

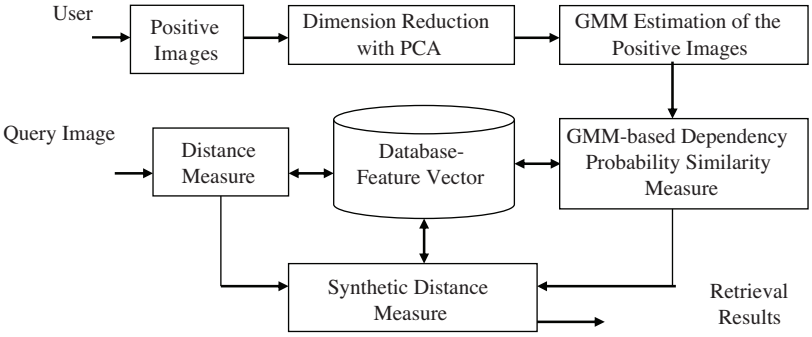

Figure 2. Block diagram of the proposed relevance feedback approach.

$$Distance^{(n+1)}(i,q) = Distance^{(n)}(i,q) - \alpha_{n+1} f(x_i|\theta_P)^{(n)}, \tag{4}$$

where $n = 0, 1, .., N$, $Distance^{(n)}(i,q)$ is the weighted Euclidean distance between the database image i and query image q in iteration n of the proposed RF, $f(x_i|\theta_P)^{(n)}$ is the GMM-based dependency probability of database image i to the distribution function of the positive images with parameters θ_P (i.e. parameters μ and Σ) that are labeled in iteration n by the user, and α_{n+1} is a constant coefficient that determines the influence of the GMM-based dependency probability measure in the synthetic distance measure. In the synthetic distance measure, $Distance^{(0)}(i,q)$ and $f(x_i|\theta_P)^{(0)}$ are the weighted Euclidean (Eq. (2)) distance between the query image and database image i and the GMM-based dependency probability of database image i to the distribution function of the positive images (Eq. (3)) in the query stage, respectively. In our distance measure, the GMM-based dependency probability (Eq. (3)) is as a similarity measure that subtracts from the distance measure until a synthetic distance measure is formed. Note that in each iteration of the proposed RF, $Distance^{(n)}(i,q)$ and $f(x_i|\theta_P)^{(n)}$ are normalized in the range of [0, 1] and then combined together.

The strategy of the determination of α_{n+1} is based on the estimation precision of the GMM in the consecutive iterations of the proposed RF, i.e. in the primary iterations of the proposed RF, the GMM (Eq. (3)) is estimated with a few images, whereas in the next iterations, the GMM estimation will be more accurate with more positive images. The distribution function estimation and the GMM-based dependency probability similarity measure are more reliable in the later iterations of the RF. Hence, the value of α_{n+1} is increased in each iteration of the proposed RF. In other words, the weights of these 2 measures in synthetic measure should be determined based on their retrieval performance in the previous iterations. Thus, in each iteration of the proposed RF, the value of α_{n+1} should correspond to the number of the retrieved positive images by 2 distance (Eq. (2)) and similarity (Eq. (2)) measures, separately. However, if the value of α_n for each query image

and each iteration of the proposed RF is determined based on ratio of the number of retrieved positive images by 2 weighted Euclidean distances (Eq. (2)) and the GMM-based dependency probability similarity (Eq. (3)) measures, then the performance of retrieval will improve. Therefore, the value of α_n is determined as:

$$\alpha_n = \frac{N_{SM}^n}{N_{DM}}, \tag{5}$$

where N_{SM}^n and N_{DM} are the number of the retrieved positive images in the nth iteration of the proposed RF by the GMM-based dependency probability similarity measure and the number of the retrieved positive images by the weighted Euclidean distance measure, respectively.

4. Experimental results

4.1. Medical X-ray image database

The database used in our work is a collection of 10,000 images consisting of 57 different radiological X-ray classes (an image sample of each class is shown in Figure 3). The images are a database of the IRMA project X-ray library (ImageCLEF, 2005) [3] that was captured and categorized by a specialist. All of the images were resized to fit into a bounding box (512 × 128 up to 512 × 512 pixels, 8 bits), where the original aspect ratio was maintained. This database has been partitioned into 2 datasets consisting of 9000 images as training datasets and 1000 images as test datasets.

Figure 3. Image sample of 57-class database.

4.2. Evaluation parameters of image retrieval

In this paper, we use the standard measures such as precision and recall to evaluate the results. These parameters are defined as below:

$$\text{Recall} = \frac{\text{Number of images retrieved and relevant}}{\text{Total number of relevent images in the database}}, \tag{6}$$

$$\text{Precision} = \frac{\text{Number of images retrieved and relevant}}{\text{Total number of retrieved images}}. \tag{7}$$

Moreover, P(R = 0.5), the precision value at the point where the recall value is 0.5; P(R = P), the precision value where the recall and precision values are equal; P-R area, the area under the P-R curve; and P(N_R), the precision after N_R images are retrieved are used in the evaluation of the previously presented systems [32].

4.3. The merging-based classification

Input images are classified by the merging-based classifier. To answer the 2 mentioned questions in Section 2.2, first, the database images are classified into 57 classes using different feature spaces until the best feature space is determined, and they are then optimized by the forward selection and GFRO algorithms, respectively. Second, homogeneous classes are created by applying the merging scheme to the classification results.

4.3.1. Feature selection, optimization, and reduction

Using the extracted features [i.e. directional histogram (D), eccentricity (E), Fourier descriptor-complex representation (FZ), invariant moments (I), major and minor axis lengths (M), major axis orientation (O), tessellation-based spectral (S) and contrast, correlation, and energy and homogeneity (T)], 13 different feature spaces are formed (Table 1). Table 1 shows the classification results for a 57-class classification problem and different feature spaces of the test dataset. Among the first 6 feature spaces of Table 1, the feature spaces with shape features, and especially the FZ feature (i.e. the 4th and 5th feature space in Table 1), obtain a higher classification accuracy rate than other feature spaces because shape features provide a considerable distinction between different classes of our database. Therefore, the last 7 feature spaces in Table 1 consist of all the extracted shape features [i.e. the feature space (FZ,I,E,M,O)]. Different feature spaces are formed with the addition of different texture features, such as contrast, correlation, energy and homogeneity, directional histogram, and tessellation-based spectral features. The highest classification accuracy rate (59.86%) is obtained by the (FZ,I,E,M,O,T,S) feature space. However, we use the GFRO algorithm for both the elimination of negligible features (weak features) and the determination of the feature weight in the optimum feature vector. We consider heuristically $T_{GA} = 0.05$ as a threshold for the detection of the weak features. The classification

Table 1. Classification results based on test dataset.

Feature space	Feature length	Accuracy % (before applying the GFRO algorithm)	Feature length (after applying the GFRO algorithm)	Accuracy % (after applying the GFRO algorithm)
M,O,T	19	24.25	9	25.23
M,O,S	19	34.66	11	36.1
M,O,D	39	27.91	18	29.18
M,O,FZ	257	43.23	112	45.67
E,M,O,FZ	258	43.84	112	46.15
E,M,O,S	20	35.11	12	38.97
FZ,I,E,M,O,D	299	53.99	154	55.2
FZ,I,E,M,O,S	279	54.81	135	56.44
FZ,I,E,M,O,T	279	54.65	138	55.46
FZ,I,E,M,O,T,D	315	54.37	160	56.26
FZ,I,E,M,O,D,S	315	58.71	158	61.56
FZ,I,E,M,O,T,S	295	59.86	152	62.19
FZ,I,E,M,O,T,S,D	331	58.27	159	63.73

performance with the optimized feature spaces has improved over those without the GFRO algorithm. The (FZ,I,E,M,O,T,S,D) feature space provides more improvement than other feature spaces. This feature space obtains a 63.73% accuracy rate for a 57-class classification problem after using the GFRO algorithm.

4.3.2. The merging scheme

In our application, using trial and error, the λ, β, and γ thresholds are set to 60%, 0.3, and 0.75, respectively. After applying the first iteration of the merging scheme, 28 merged classes are obtained, while the total classification accuracy rate is improved to 88.9%. Because the total accuracy rate is lower than the desired value ($T_{desired} = 90\%$), the second iteration of the merging scheme is executed. After applying the second iteration of the merging scheme, the total classification accuracy rate is improved to 90.23% for a 23-class classification problem (Table 2). In this iteration, the desired value of the total accuracy rate has been obtained, and so the merging scheme is terminated.

Table 2. Results of the merging scheme in the second iteration.

The merged class after applying the marging scheme	Class name	Classification accuracy (%) of the merged class	The merged class after applying the merging scheme	Class name	Classification accuracy (%) of the merged class
C1	1, 45	95.17	C13	32, 35	86
C2	3	93.33	C14	41	92
C3	4	91.25	C15	42	90
C4	5	96.77	C16	43	91
C5	6, 7, 8, 9, 18, 31, 36, 37, 19, 24, 20, 29, 47	91	C17	48	94
C6	10, 14, 34, 15, 23, 26, 46, 51	89.9	C18	49	95
C7	2, 11, 12, 13, 16, 33, 40, 44	99	C19	50	73
C8	17	94.66	C20	52	95
C9	21, 22, 30, 38, 39	89	C21	54, 56	91
C10	25	93.45	C22	55	88
C11	27, 53	78.36	C23	57	98
C12	28	88			

4.4. Search space and similarity measure selections

According to the block diagram in Figure 1, the m-nearest classes to the query image are determined as the search space for image retrieval. The selection of m and the optimum distance measure are 2 important parameters in the performance of our proposed CBIR system. The selection of m is a compromise between the speed and the performance of the retrieval, i.e. if m is set to higher values, the existence probability of the relevant images in the search space is increased, but the retrieval speed is decreased with the extension of the search space and vice versa. The classification results for different values of m are shown in Table 3. We set m to 5 since this value guarantees the existence of relevant images in 99.66% of the query images.

In CBIR systems, a similarity measure is applied to sort the images based on their similarity. In our framework, the similarity measure of 2 images is defined based on the distance of their feature vectors in the feature space. After evaluating several measures, such as correlation, city block, cosine, sEuclidean, Chebyshev, and weighted Euclidean (Table 4), we selected the weighted Euclidean distance (Eq. (2)) as the dissimilarity measure in our image retrieval framework.

Table 3. Classification results in the m-class of the closest classes to the query image.

m-Class of the closest classes	Accuracy rate (%)
1	90.23
2	96.34
3	98.18
4	99.35
5	99.66
6	99.72

Table 4. Retrieval results for different distance measures.

Distance measure	Retrieval measures	
	P(P = R)	P(20)
Correlation	0.29	0.36
Weighted Euclidean	0.38	0.48
City block	0.32	0.39
sEuclidean	0.34	0.4
Cosine	0.22	0.35
Chebyshev	0.33	0.45

4.5. Interactive retrieval with the proposed RF

To set up the proposed RF properly, the determination of 2 parameters is critical. The first is the optimal number of Gaussian functions in the GMM estimation of the GMM-based dependency probability measure (Eq. (3)) and the second is the value of coefficient α_n in the synthetic distance measure in each iteration of the proposed RF.

To determine the optimum number of Gaussian functions in the GMM estimation, several experiments are designed as follows. The RF is implemented based on the GMM-based dependency probability of the database images into the GMM of the positive images, as a similarity measure (similar to the RF method presented in [30]), whereas the number of Gaussian functions (Z) in the GMM estimation is varied. Retrieval results based on the GMM-based dependency probability similarity measure (Eq. (3)) are presented in Table 5.

Table 5. Retrieval results based on the GMM-based dependency probability similarity measure.

# Gaussian functions (Z)	P(P = R)			P(20)		
	1st RF	2nd RF	3rd RF	1st RF	2nd RF	3rd RF
2	0.31	0.40	0.42	0.46	0.71	0.74
3	0.32	0.41	0.43	0.47	0.75	0.78
4	0.35	0.43	0.46	0.49	0.78	0.79
5	0.36	0.46	0.49	0.49	0.79	0.81
6	0.36	0.46	0.49	0.49	0.8	0.8
Without RF	0.38			0.48		

By evaluation of the retrieval results in Table 5, we infer that the GMM estimation with 5 Gaussian functions ($Z = 5$) is appropriate. The retrieval result in the first iteration of the RF is worse than the retrieval without RF (Table 5). P(P = R) values in the retrieval without RF and the first iteration of the RF are 0.38 and 0.36, respectively, which shows a negligible decrease in the retrieval performance. However, in the next iterations of the RF, the retrieval performance improves considerably.

The second important parameter in our proposed RF is coefficient α_n. The retrieval results in the first 3 iterations of the RF for different values of α_n are shown in Table 6. The retrieval results show that if the value of α_n is gradually increased in the consecutive iterations of the RF, better results are obtained. This is due to a more accurate estimation of the positive image distribution function in the consecutive iterations of the RF. However, if the value of α_n is calculated using Eq. (5), the improvement of the retrieved performance will increase considerably.

Table 6. Retrieval results with our proposed RF for different values of α based on the test dataset.

Coefficient of α			P(P = R)			P(20)		
α_1	α_2	α_3	1st RF	2nd RF	3rd RF	1st RF	2nd RF	3rd RF
1	1	1	0.49	0.58	0.61	0.85	0.86	0.86
0.5	0.7	0.9	0.64	0.65	0.67	0.84	0.85	0.86
0.9	0.7	0.5	0.51	0.59	0.61	0.81	0.83	0.84
According to Eq. (6)			0.67	0.69	0.71	0.91	0.915	0.915
Without RF			0.38			0.48		

Figure 4 shows the precision-recall curve for the image retrieval without the RF and the first 3 iterations of the RF based on the test dataset when Z is set to 5 and α_n is determined by Eq. (5). The retrieval results in the first 3 iterations of the RF obtain a P-R area of 0.63, 0.65, and 0.67 and P(R = 0.5) of 0.85, 0.87, and 0.88, respectively (precision-recall curves shown in Figure 4).

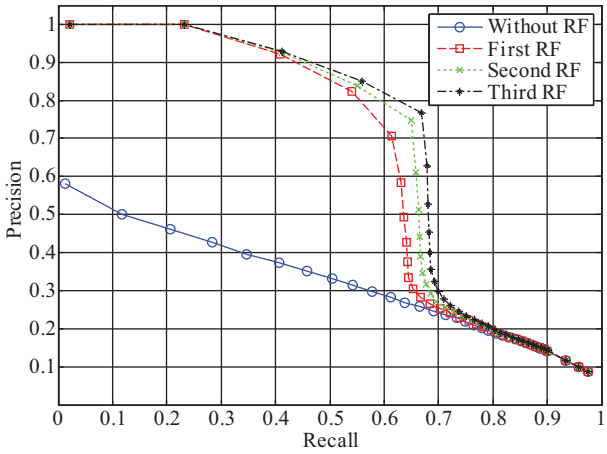

Figure 4. Precision-recall curves for image retrieval without RF and the first 3 iterations of the proposed RF.

5. Discussion

The retrieval results of our proposed framework were obtained based on the ImageCLEF 2005 dataset, consisting of 9000 images as a training dataset and 1000 images as a test dataset. The performance of this framework was evaluated based on 4 criteria, P(20), P(R = P), P(R = 0.5), and P-R area, where the last 3 criteria were extracted from the precision-recall curve. In the third iteration of the RF, 0.915, 0.71, 0.88, and 0.67 were achieved for P(20), P(R = P), P(R = 0.5), and P-R area, respectively.

Due to the lack of a standard dataset, a perfect and precise comparison between the presented algorithms in the literature and our proposed algorithm is a complex task [16]. However, to evaluate the results of our proposed algorithm, we obtain a comparison between our proposed algorithm and other presented retrieval

techniques in the literature. Several retrieval algorithms can be discussed. The best reported retrieval result in [16] with a dataset of 1501 radiological images of 17 classes was 0.67, 0.62, and 0.66 for P(R = 0.5), P(R = P), and P-R area, respectively (these values had been approximately calculated from the precision-recall curve in [16]). Developing the presented GMM-KL framework in [16] for a large dataset is a problem, particularly due to the computational complexity of the KL measure calculation. However, applying this extension to a larger database in our proposed CBIR framework is not a challenge. The best performance for the proposed CBIR system in [17], based on a database consisting of 5000 images of 20 classes, was 0.82, 0.68, and 0.72 for P(R = 0.5), P(R = P), and P-R area, respectively (these values had been approximately calculated from the precision-recall curve in [17]). In this system, the retrieval precision never reached 0.9, whereas the precision of our proposed CBIR framework was equal to 1 in the first 15 images of the retrieved images. However, the algorithm in [17] obtained better retrieval results when more images were retrieved because the value of the P-R area measure (i.e. 0.72) in [17] was greater than that of our proposed framework (i.e. 0.67). The presented classification-based image retrieval framework in [18] was evaluated on the ImageCLEFmed'06 database [33] (consisting of 10,000 images as the training dataset and 1000 images as the test dataset). The best reported retrieval results were 0.64, 0.61, and 0.58 for P(R = 0.5), P(R = P), and P-R area, respectively (these values had been approximately calculated from the precision-recall curve in [18]). The database used in [18] was larger than our database. Hence, the improvement of our results is predictable. A summary of these comparisons are presented in Table 7.

Table 7. Comparison between the proposed RF and the previous works (see text).

Approach	Retrieval measures		
	P(R = P)	P(R = 0.5)	P-R area
Algorithm [16] (without RF)	0.62	0.67	0.66
Algorithm [17] (with RF)	0.68	0.82	0.72
Algorithm [18] (without RF)	0.61	0.64	0.58
Proposed algorithm (with RF)	0.71	0.88	0.67

6. Conclusion

In this paper, a content-based medical X-ray image retrieval framework was presented. This system was designed based on a merging-based classification algorithm and the weighted Euclidean distance measure for image retrieval in a large database. The merging-based classification step could reduce both the computational complexity and the false acceptance rate of the CBIR system. Using the merging scheme and the proposed GFRO algorithm, not only were the homogenous classes formed, but the classification performance was also considerably improved. A novel synthetic RF approach was integrated into our proposed image retrieval framework for the improvement of the retrieval performance and to narrow down the semantic gap. This RF approach was based on the combination of the weighted Euclidean distance and the GMM-based dependency probability similarity measures. Our proposed CBIR framework with RF was evaluated on a database consisting of 10,000 medical X-ray images of 57 predefined classes. Analysis of the retrieval results based on the precision-recall curves without and with the RF was carried out with 1000 query images. The effectiveness of the proposed framework compared with other presented retrieval algorithms in the literature was demonstrated by our obtained results.

Acknowledgment

The authors would like to thank the IRMA Group, Aachen, Germany, for making the database available for the experiments.

References

[1] H. Pourghassem, H. Ghassemian, "Content-based medical image classification using a new hierarchical merging scheme", Journal of Computerized Medical Imaging and Graphics, Vol. 22, pp. 651–661, 2008.

[2] C. Shyu, C. Brodley, A. Kak, A. Kosaka, A. Aisen, L. Broderick, "ASSERT: A physician-in-the-loop content-based image retrieval system for HRCT image databases", Computer Vision and Image Understanding, Vol. 75, pp. 111–132, 1999.

[3] T. Lehmann, M. Guld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, B.B. Wein, "Content-based image retrieval in medical applications", Methods of Information in Medicine, Vol. 43, pp. 354–361, 2004.

[4] S. Antani, D.J. Lee, L.R. Long, G.R. Thoma, "Evaluation of shape similarity measurement methods for spine X-ray images", Journal of Visual Communication and Image Representation, Vol. 15, pp. 285–302, 2004.

[5] J.G. Dy, C.E. Brodley, A. Kak, L.S. Broderick, A.M. Aisen, "Unsupervised feature selection applied to content-based retrieval of lung images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, pp. 373–378, 2003.

[6] C.R. Shyu, C.A. Pavlopoulou, C. Kak, C.E. Brodley, "Using human perceptual categories for content-based retrieval from a medical image database", Computer Vision and Image Understanding, Vol. 88, pp. 119–151, 2002.

[7] P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, Z. Protopapas, "Fast and effective retrieval of medical tumor shapes", IEEE Transactions on Knowledge and Data Engineering, Vol. 10, pp. 889–904, 1998.

[8] S.N. Yu, C.T. Chianga, C.C. Hsieh, "A three-object model for the similarity searches of chest CT images", Computerized Medical Imaging and Graphics, Vol. 29, pp. 617–630, 2005.

[9] L.L.G. Oliveira, S.A. Silva, L.H.V. Ribeiro, R.M. Oliveira, C. Coelho, A.S.S. Andrade, "Computer-aided diagnosis in chest radiography for detection of childhood pneumonia", International Journal of Medical Informatics, Vol. 77, pp. 555–564, 2007.

[10] L.R. Long, S. Antania, D.J. Leeb, D.M. Krainakc, G.R. Thoma, "Biomedical information from a national collection of spine X-rays film to content-based retrieval" Proceedings of SPIE, Vol. 5033, pp. 70–84, 2003.

[11] L.R. Long, K. Sameer, A. George, R. Thoma, "Image informatics at a national research center", Computerized Medical Imaging and Graphics, Vol. 29, pp. 171–193, 2005.

[12] X. Xu, D.J. Lee, S. Antani, L.R. Long, "A spine X-ray image retrieval system using partial shape matching", IEEE Transactions on Information Technology in Biomedicine, Vol. 12, pp. 100–108, 2008.

[13] O. Nomira, M. Abdel-Mottalebb, "Hierarchical contour matching for dental X-ray radiographs", Pattern Recognition, Vol. 41, pp. 130–138, 2008.

[14] H. Müller, A. Rosset, J.P. Vallée, A. Geissbuhler, "Comparing feature sets for content-based medical information retrieval", Proceedings of SPIE Medical Imaging, Vol. 5351, pp. 99–109, 2004.

[15] W.W. Chu, C.C. Hsu, A.F. Cardenas, R.K. Taira, "Knowledge-based image retrieval with spatial and temporal constructs", IEEE Transactions on Knowledge and Data Engineering, Vol. 10, pp. 872–888, 1998.

[16] H. Greenspan, A.T. Pinhas, "Medical image categorization and retrieval for PACS using the GMM-KL framework", IEEE Transactions on Information Technology in Biomedicine, Vol. 11, pp. 190–202, 2007.

[17] M.M. Rahman, P. Bhattacharya, B.C. Desai, "A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback", IEEE Transactions on Information Technology in Biomedicine, Vol. 11, pp. 58–69, 2007.

[18] M.M. Rahman, B.C. Desai, P. Bhattacharya, "Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion", Computerized Medical Imaging and Graphics, Vol. 32, pp. 95–108, 2008.

[19] J. Yao, Z. Zhang, S. Antani, R. Long, G. Thoma, "Automatic medical image annotation and retrieval", Neurocomputing, Vol. 71, pp. 2012–2022, 2008.

[20] L. Yang, F. Algregtsen, "Fast computation of invariant geometric moments: a new method giving correct results", Proceedings of the 12th IAPR International Conference on Pattern Recognition. Conference A: Computer Vision & Image Processing, Vol. 1, pp. 201–204, 1994.

[21] N. Otsu, "A threshold selection method from gray-scale histogram," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 8, pp. 62–66, 1978.

[22] E. Persoon, K. Fu, "Shape discrimination using Fourier descriptors", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 7, pp. 170–79, 1977.

[23] A.K. Jain, Fundamentals of Digital Image Processing, New Jersey, Prentice Hall, 1989.

[24] R.M. Haralick, K. Shanmugan, I. Dinstein, "Textural features for image classification", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 3, pp. 610–621, 1973.

[25] A.L. Blum, P. Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, Vol. 97, pp. 245–271, 1997.

[26] S.X. Yu, "Feature selection and classifier ensembles: a study on hyperspectral remote sensing data", PhD, University of Antwerp, 2005.

[27] M. Mitchell, An Introduction to Genetic Algorithms, Cambridge, MA, MIT Press, 1996.

[28] E. de Ves, J. Domingo, G. Ayala, P. Zuccarello, "A novel Bayesian framework for relevance feedback in image content-based retrieval systems", Pattern Recognition, Vol. 39, pp. 1622–1632, 2006.

[29] P.C. Cheng, B.C. Chien, H.R. Ke, W.P. Yang, "A two-level relevance feedback mechanism for image retrieval", Expert Systems with Applications, Vol. 34, pp. 2193–2200, 2008.

[30] F. Qian, M. Li, L. Zhang, H. Zhang, B. Zhang, "Gaussian mixture model for relevance feedback in image retrieval", IEEE International Conference on Multimedia and Expo, Vol. 1, pp. 229–232, 2002.

[31] A.K. Jain, B. Bhandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice", in: P.R. Krishnaiah, L.N. Kanal, Eds., Handbook of Statistics, Amsterdam, Elsevier, pp. 835–855, 1987.

[32] T. Deselaers, D. Keysers, H. Ney, "Classification error rate for quantitative evaluation of content-based image retrieval systems", Proceedings of the 17th International Conference on Pattern Recognition, Vol. 2, pp. 505–508, 2004.

[33] H. Muller, T. Deselaers, T.M. Lehmann, P. Clough, E. Kim, W. Hersh, "Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks", CLEF Proceedings, Lecture Notes in Computer Science, pp. 595–608, 2007.