J Digit Imaging (2014) 27:77–89
DOI 10.1007/s10278-013-9637-0

Automatic Medical X-ray Image Classification using Annotation

Mohammad Reza Zare · Ahmed Mueen · Woo Chaw Seng

Published online: 3 October 2013

Abstract The demand for automatically classification of medical X-ray images is rising faster than ever. In this paper, an approach is presented to gain high accuracy rate for those classes of medical database with high ratio of intraclass variability and interclass similarities. The classification framework was constructed via annotation using the following three techniques: annotation by binary classification, annotation by probabilistic latent semantic analysis, and annotation using top similar images. Next, final annotation was constructed by applying ranking similarity on annotated keywords made by each technique. The final annotation keywords were then divided into three levels according to the body region, specific bone structure in body region as well as imaging direction. Different weights were given to each level of the keywords; they are then used to calculate the weightage for each category of medical images based on their ground truth annotation. The weightage computed from the generated annotation of query image was compared with the weightage of each category of medical images, and then the query image would be assigned to the category with closest weightage to the query image. The average accuracy rate reported is 87.5 %.

Keywords Image classification · Medical X-ray images · BoW · PLSA · SVM

M. R. Zare (✉) · W. C. Seng
Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
e-mail: mreza_zare57@yahoo.com

W. C. Seng
e-mail: cswoo@um.edu.my

A. Mueen
Department of Computer and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: mueen@kau.edu.sa

Background

Over the last decade, storage of non-text-based data in the database has become an increasingly important trend in information management. Many medical images are acquired everyday in any modern hospital due to the rapid development of digital medical imaging techniques and information technologies. As a result, there is an increased demand for a computerized system to manage these valuable resources.

Currently, many hospitals and radiography departments are equipped with picture archiving and communications system. Such traditional systems have many limitations due to the usage of Digital Imaging and Communications in Medicine (DICOM) header as the searches are carried out according to the textual attributes of image headers. Even though DICOM header contains many important information, it still remains suboptimal due to its high error rate reported in recent studies [1].

Content-based image retrieval (CBIR) enables the elimination of such difficulties that exist in those traditional systems where the searches only performed based on DICOM header [2]. CBIR deals with the analysis of image content and the development of tools to represent visual content in a way that can be efficiently searched and compared. The objective of any CBIR system is to retrieve the similar images to the query image in the most efficient and effective way. In the medical domain, such retrieval system can also provide diagnostic support to physicians or radiologists by displaying relevant past cases to assist them in decision-making process. Besides diagnostics, medical image retrieval can also be useful in education and research by providing visual access in existing large repositories.

It is believed that the quality of such medical system can be improved by a successful classification of images by filtering out irrelevant images. For instance, the process to search images for a query like “Find Anteroposterior (AP) Lumbar

Spine X-ray image" starts with pre-filtering the database images according to the imaging modality (X-ray), body region (spine), and orientation (AP). Then, the search could be performed on the set of filtered images to find specific sub-body region such as the "lumbar spine." However, unlike earlier years of this research that the classification of medical images was restricted to few classes only, this task is challenged when it deals with large archive medical database.

This is where the ImageCLEF medical image annotation challenge was born. The goal of this challenge is to classify the images into pre-defined classes automatically and assigning the correct labels to unseen test images. The database used in this study is ImageCLEF 2007 [3] which was provided by the IRMA group from RWTH University Hospital of Aachen, Germany. It consists of medical radiographs collected randomly from daily routine work at the Department of Diagnostic Radiology. The quality of radiographs varies considerably, and there is a high intraclass variability and interclass similarity among classes. In order to establish a ground truth, the images were manually classified by expert physicians using the IRMA code [4]. The four main facets of IRMA code are image modality also known as technical (T), body orientation known as directional (D), body region examined also called anatomical (A), and biological system called biological (B). Sample images from the database together with textual labels and their complete code are given in Fig. 1.

The classification task begins with extracting appropriate visual features of the image. It is one of the most important factors in design process of such system. Moreover, feature extraction step affects all other subsequent processes. Visual features were categorized into primitive features such as color, shape, and texture. However, as X-ray images are gray level images and do not contain any color information, the related CBIR systems mostly deal with textures for feature extraction process which were used by several researchers [5–17]. Gray level co-occurrence matrices (GLCM) [18] and local binary patterns introduced by Ojala et al. in [19] are commonly used feature extraction techniques in the above works.

Hierarchical classification schemes based on individual SVMs trained on IRMA sub-codes for the task of automatic annotation of ImageCLEF 2007 medical database was proposed in [17]. Another widely used strategy is combining different local and global descriptors into a unique feature representation. A combination of multi-visual features such as GLCM, pixel value, and canney edge detector as shape feature was presented by Mueen et al. [8]. The accuracy rate obtained by their algorithm on ImageCLEF 2005 with 57 classes was 89 %. Other authors have also combined pixel value as a global image descriptor with other image representation techniques to construct feature vector of the image [9, 14]. The accuracy rate reported by the method proposed in [9] and [14] are 89.7 and 81.96 %, respectively.

Recently, more promising studies have been focused on local patch-based image representation. The bag of words (BoW) represents images using histograms of quantized appearances of local patches. Such methods are constructed by extracting features around the interest points of the image. In recent years, many studies have successfully exploited this feature in general scene and object recognition tasks [20–24]. The use of BoW model can also be found in medical image classification and retrieval tasks [9, 14, 25–31].

With increasing size of medical X-ray archives, it is important to have simplistic, discrete representations and simple matching measures to preserve computational efficiency. It is argued that BoW paradigm provides efficient means to address the challenge of CBIR system in large size databases such as the one in ImageCLEF [27]. Avni et al. in [27] proposed X-ray image categorization and retrieval based on local patch representations using BoW approach. This was an extension of another work where visual words dictionary were generated to represent X-Ray chest images [26].

The combination of local and global features was used to address the problem of intraclass variability and interclass similarity for the task of medical image classification in [9]. They integrated two different local cues that describe structural and textual information of image patches. They reported an accuracy rate of 89.7 % on ImageCLEF 2007 database.

However, the results obtained in all of the above presented works are at a global level, meaning the performance is obtained on the entire database. By observing and analyzing the accuracy rate of every individual class, it is clear that almost all large categories of medical images in the database have accuracy rates of above 85 % whereas images from small

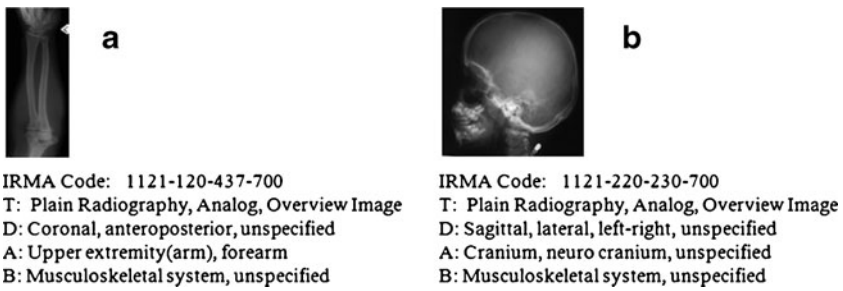

Fig. 1 Sample images and their corresponding IRMA code

classes are frequently misclassified. This observation shows that discriminatively trained classifiers are usually more accurate when labeled training data is abundant. At the same time, most of the classes with low accuracy rate are those with high ratio of intraclass variability and interclass similarity. To address this issue and increase the number of classes with high accuracy rate, a classification framework was developed by the authors of this study [30] to perform filtering on the dataset in several iterations, and consequently, a separate model is constructed from each iteration. The idea is to filter out classes with good accuracy rate in the first iteration. Subsequently, the next iteration only deals with less predominant classes. Indeed, the generated model constructed in every iteration consists of those classes with an optimum accuracy rate. The threshold of 80 % has been set for the optimum accuracy rate. This threshold is chosen because it is very rare to have high percentage of accuracy in large medical database. We had chosen a balanced value here to trade off accuracy with practicality.

After the first iteration, there were about 39 classes with accuracy below 80 %. These classes were combined and gone through the classification process again to create another classification model during the second iteration. The accuracy rate reported by the model constructed from this iteration was 72 % even though this model was constructed on smaller number classes as compared to the model constructed from the first iteration. In this study, we analyze the classification performance of the model generated from the second iteration and propose a technique to improve the accuracy rate. Detailed analysis has been done on classification results of every individual class to find out the reason of low accuracy rate. The analysis showed that most of them are misclassified within their own subregion. In another word, those images that are misclassified are visually similar; this is due to the high ratio of intraclass variability and interclass similarities. We take an example of one of the "arm" sub-body region which is the "forearm." There are four classes under this sub-body region; the images in these four classes are visually similar. What makes these four classes to be distinguished from one another is imaging view and direction. This examination shows that depending on only one technique to gain high accuracy rate for every individual class of such database with the said complexity is unreliable. The fact that most of the misclassifications is from those classes with the same body region motivated us to do the classification task by using different annotation techniques. Annotation performance of every technique varies from one another depending on the body region in medical database. The idea is to take advantage of different approaches in annotation to get to the closest class/category to the test image. Indeed, each annotation technique is complementary for the other annotation techniques. As a result, the combined set of keywords generated from these annotation techniques would represent the image category clearly.

One of the widely used approaches for annotation task is based on classification. This approach has been applied to a number of image classification and annotation tasks [5, 8, 17].

Apart from the classification approach, another widely used annotation technique annotates an image with multiple semantic concepts/categories [32–34]. This approach takes a different stand and treats image as collection of visual words or BoW. One of the Bayesian model such as probabilistic latent semantic analysis (PLSA) [35] works with BoW features and has been successfully applied to annotate and retrieve images [32–34, 36]. In [33], authors presented a semantic annotation model which employs continuous PLSA and standard PLSA to model visual features and textual words, respectively. The model learns the correlation between these two modalities by an asymmetric learning approach, and then it can predict semantic annotation for unseen images. Multi-modal probabilistic latent semantic analysis which incorporates visual features and tags by generating semantic contexts was proposed in [34]. Zare et al. [37] have also employed PLSA to generate a robust, high level representation and low-dimensional image representation of medical images in order to disambiguate the BoW representation.

In this paper, both of the above approaches of automatic annotation are incorporated for an automatic classification of medical X-ray images. The rest of the paper is organized as follows: "Bag of words image representation" discusses the BoW representation of images. "PLSA Model" presents the key concept of the PLSA model. The proposed classification approach is discussed in "Materials and methods". Experimental results and discussion are reported and analyzed in "Experimental results" and "Discussion," respectively. Finally, the overall conclusion of this study is presented in "Conclusion."

Bag of Words Image Representation

The process of BoW started with detecting local interest point. Local interest point detectors have the task of extracting specific points and areas from images which are invariant to some geometric and photometric transformations. One of the popular approaches for the detection of local interest point is difference of Gaussians (DoG) which is used in this experiment. This detector has been chosen since it was shown to perform well for the task of wide-baseline matching when compared to other detectors. We can observe that the DoG detector is considerably faster since it is based on the subtraction of images. DoG has been built to be invariant to translation, scale, rotation, and illumination changes and samples images at different locations and scales. This technique uses scale-space peaks in the difference of Gaussian operator convolved with the image. Next, the detected keypoints are then represented using scale

invariant feature transform (SIFT) [38]. In short, the image gradient is sampled and its orientation is quantized. Using a grid division of the local interest area, local gradient orientation histograms are created where the gradient magnitude is accumulated. The final feature is the concatenation of all the local gradient orientation histograms. A Gaussian weighting is introduce in the SIFT feature extraction process to give more importance to samples closer to the center of the local interest area. This contributes to a greater invariance of the SIFT descriptor since samples closer to the center of the local interest areas are more robust to errors in the local interest area estimation.

In the study of Lowe [39], it was found that the best compromise between performance and speed was obtained by using a 16×16 gradient sampling grid and a 4×4 subhistogram grouping. The final descriptor proposed in this formulation is eight orientations and 4×4 blocks, resulting in a descriptor of 128 dimensions. Next step in implementation of bag of visual words is the codebook construction where the 128-dimensional local image features have to be quantized into discrete visual words. This task is performed using clustering or vector quantization algorithm. This step usually uses k-means clustering method, which clusters the keypoint descriptors in their feature space into a large number of clusters and encodes each keypoint by the index of the cluster to which it belongs. We conceive each cluster as a visual word that represents a specific local pattern shared by the keypoints in that cluster. Thus, the clustering process generates a visual word vocabulary describing different local patterns in images. The number of clusters determines the size of the vocabulary, which can vary from hundreds to over tens of thousands. Mapping the keypoints to visual words, we can represent each image as a "bag of visual words."

PLSA Model

The PLSA was originally proposed by Hoffman [35] in the context of text document retrieval. It is used to discover topics in a document using the bag of words document representations. It has also been applied to various computer vision problems such as classification, images retrieval, where we have images as documents and the discovered topics are object categories (e.g., airplane and sky). In this section, PLSA model explained in terms of images, visual words, and topics.

The key concept of the PLSA model is to map the high dimensional word distribution vector of an image to a lower dimensional topic vector. Therefore, PLSA introduces a topic layer between images and words. Suppose we have a set of images D=$d_1,\ldots, d_N$ with words from visual vocabulary X. Each image consists of mixture of multiple topics, and thus, the occurrence of words is a result of the topic mixture. PLSA assumes the existence of a latent aspect z_k ($k \in 1,\ldots,N_z$) in a generative process of each word x_j ($j \in 1,\ldots,N_x$) in the image d_i ($i \in 1,\ldots,N_d$). Each occurrence x_j is independent from the image it belongs to given the latent variable z_k, which corresponds to the joint probability expressed by

$$P(x_j, z_k, d_i) = P(d_i)P(z_k|d_i)P(x_j|z_k) \quad (1)$$

The joint probability of the observed variables is the marginalization over the N_z latent aspects z_k as expressed by

$$P(x_j, d_i) = P(d_i)\sum_{k=1}^{N_z} P\left(z_k\middle|d_i\right)P\left(x_j\middle|z_k\right) \quad (2)$$

The unobservable probability distribution $P(z_k|d_i)$ and $P(x_j|z_k)$ are learned from the data using the expectation–maximization (EM) algorithm. $P(z_k|d_i)$ denotes the probability of topic z_k given in image d_i.$P(x_j|z_k)$ denotes the probability of visual word x_j in topic z_k EM algorithm is a standard iterative technique for maximum likelihood estimation in latent variable models such as Log likelihood. Normally, 100–150 iterations are needed before data convergence.

Each iteration is composed of two steps:

1. An expectation (E) step where, based on the current estimates of the parameters, posterior probabilities are computed for the latent variable z_k.
2. A maximization (M) step, where parameters are updated for given posterior probabilities computed in the previous E step. It increases the likelihood in every step and converges to a maximum of the likelihood.

Materials and Methods

Classification process is in two steps of training and testing phases. In training phase, the selected features are extracted from all the training images and classifier is trained on the extracted features to create a model. This model is then used in testing phase to classify the unseen test image into one of the pre-defined categories. As stated earlier, the purpose of this study is to improve the classification performance of the model generated from the classes left with low accuracy rate in second iteration of the previous studies [30]. As such, we proposed to classify the unseen test image via three techniques of annotation as described below.

Annotation Using Classification

In this approach, supervised learning approach is used to classify images. Classification process consists of two steps of training and testing phases. In the training phase, the selected features are extracted from all the training images and a classifier is trained on the extracted features to create a

model as described in previous section. This model is used to classify the query images into a pre-defined class, and then corresponding keywords of that class will be assigned to the query image as an annotation. For instance, if an unseen test image classifies to class 23, it will be annotated by the following keywords: arm, forearm, wrist joint, elbow joint, radius, ulna, left, and AP view.

Annotation Using PLSA

In formulation of annotation using PLSA, we incorporate both visual vocabulary and textual vocabulary in construction of multi-modal PLSA model. Each modality (visual vocabulary and textual vocabulary) are treated differently. Based on our empirical studies, we give more importance to textual vocabulary in order to capture meaningful aspects in the data and use them for annotation. This is to ensure the consistent set of textual words is predicted while retaining the ability to jointly model the visual features. To formulate visual vocabulary, we computed a co-occurrence table where an image is represented by BoW as explained in previous section. The BoW is represented as two-dimensional matrix with 564 rows and 500 columns. Five hundred sixty-four and 500 are the numbers of training images and visual vocabulary size, respectively. The process of constructing textual vocabulary is described below.

Textual Vocabulary

Based on given IRMA code and comments given by qualified physician, the corresponding annotated keywords for each class of the medical database are identified. After eliminating the duplicate keywords, the unique set of annotated keywords are generated. An average of five keywords is specified for every image in the class. For instance, the annotated keywords that were assigned to Fig. 1a are: "Arm, Forearm, Wrist Joint, Elbow Joint, Ulna, AP View". "Arm, Forearm, AP View" were taken from the textual labels come with IRMA code; "Wrist Joint, Elbow Joint, Ulna" were given by the physician.

To formulate the textual vocabulary, the dataset is then represented as term–document matrix as shown in Fig. 2 by placing the image names and generated keywords in rows captions and columns captions of the matrix, respectively. Each cell of the matrix is then filled by 1 or 0, where 1 represents the occurrence of the particular keyword and 0 indicates the nonoccurrence of that keyword for a specific image. The final term–document matrix is represented as two-dimensional matrixes with 564 rows and 68 columns. Five hundred sixty-four and 68 are the numbers of training images and number of textual words, respectively.

Automatic Image Annotation with PLSA

Upon construction of BoW and textual vocabulary, linked pair of PLSA models is trained for the task of automatic image annotation as described below. The flow of learning and annotation on the unseen test image is illustrated in Fig. 3.

Learning Phase

1. First, PLSA model is completely trained on the set of image captions (textual vocabulary) to learn both $P(x_j|z_k)$ and $P(z_k|d_i)$ parameters. As a result, set of aspects automatically learned on the textual vocabularies and their most probable training images.
2. We then consider that the aspects have been observed for these set of images **d** and train a second PLSA on the visual modality (BoW) to compute $P(x_j|z_k)$, keeping $P(z_k|d_i)$ fix which was learnt from step one. The resulting value for $P(x_j|z_k)$ is presented as two-dimensional matrix XZ, where X is the number of textual words and Z is the number of classes in dataset. In this experiment, the value of textual words (X) and number of classes (Z) is 68 and 39, respectively.

Automatic Annotation on the Test Image

1. Given new visual features from the unseen test image and the previously calculated $P(x_j|z_k)$ parameters, $P(z_k|d_{\text{test}})$

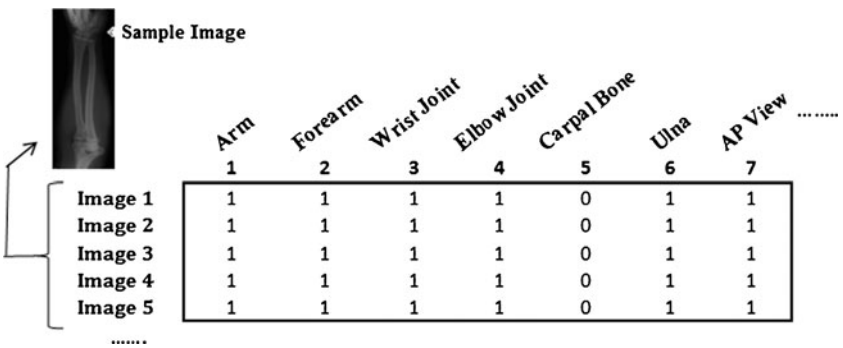

Fig. 2 Sample term–document matrix of textual vocabulary

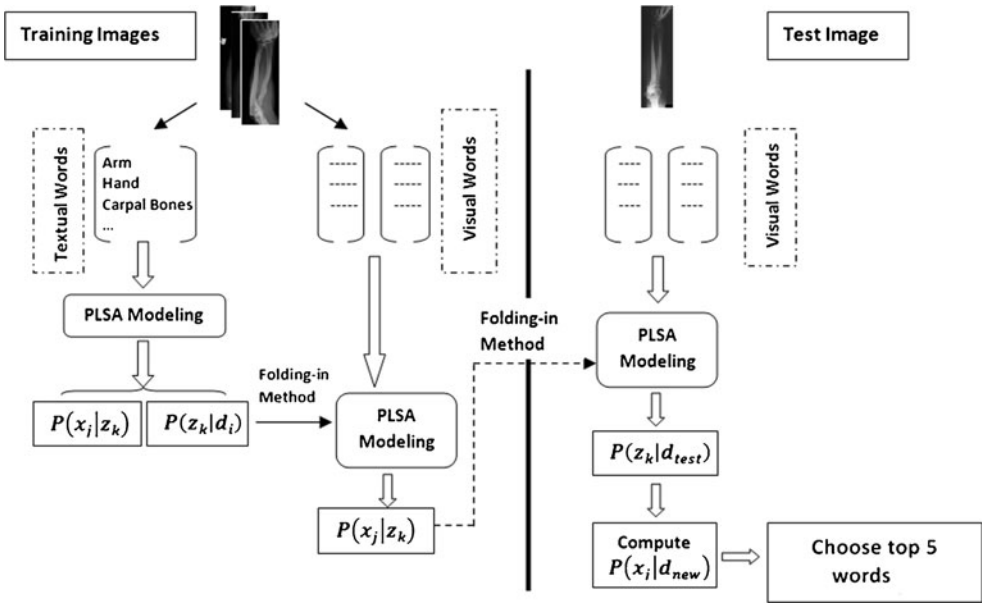

Fig. 3 The flow of training and annotating using PLSA

is computed for a new image d_{new} using the standard PLSA procedure for a new document. Similar to $P(x_j|z_k)$, the resulting value for $P(z_k|d_{\text{test}})$ is shown as two-dimensional matrix ZD where D (e.x. 5) is the number of test images.

2. The posterior probability of each word in the vocabulary is then computed by

$$P\left(x_j\middle|d_{\text{new}}\right) = \sum_{k=1}^{k} P\left(z_k\middle|d_{\text{new}}\right) P\left(x_j\middle|z_k\right) \tag{3}$$

This is performed by multiplying the two matrixes as follows:

$$C = XZ \times ZD$$

The result of multiplication is a two-dimensional matrix C with 68 rows and five columns where each column represents one of the test images. Thus, the top highest five numbers in every column are chosen where each one of them represents a word. As a result, the number of annotated keywords is five.

Annotation Using Top Similar Images

In this approach, the top five training images that are visually similar to the query image would be selected followed by identifying the class that they belong to. The corresponding keywords to each class would be then taken as an annotation. These five sets of keywords are then combined to produce distinct set of keywords. The block diagram of retrieving similar images using PLSA is shown in Fig. 4. The process of retrieving the top five similar training images is as follows:

Learning Phase

1. A first PLSA model is completely trained on the set of training images with visual words (BoW) as an input to learn both $P(z_k|d_i)$ and $P(x_j|z_k)$.
2. While not converge do
   a. E step: Compute the posterior probabilities $P(z_k|d_i,x_j)$
   b. M step: Parameters $P(x_j|z_k)$ and $P(z_k|d_i)$ are updated from posterior probabilities computed in the E step.

End While

Testing Phase

1. The E step and M step are applied on the extracted BoW of the test image by keeping the probability of $P(x_j|z_k)$ learnt from the training phase fixed.
2. Calculate the Euclidean distance between $P(z_k|d_i)$ and $(z_k|d_{\text{test}})$.
3. Those images with closest distance to $P(z_k|d_{\text{test}})$ will be retrieved as similar images.

Applying Ranking Similarity to Produce Final Annotation

Based on the proposed algorithm, three sets of keywords were generated based on the above three approaches. Each set of the generated keywords was ranked according to their importance to describe the image. Two levels of ranking are applied on the keywords; those keywords which help to distinguish the body region clearly were ranked as the first level and those that describe the objects (specific bone structure) inside the specific body region as well as imaging direction and view are ranked as second level.

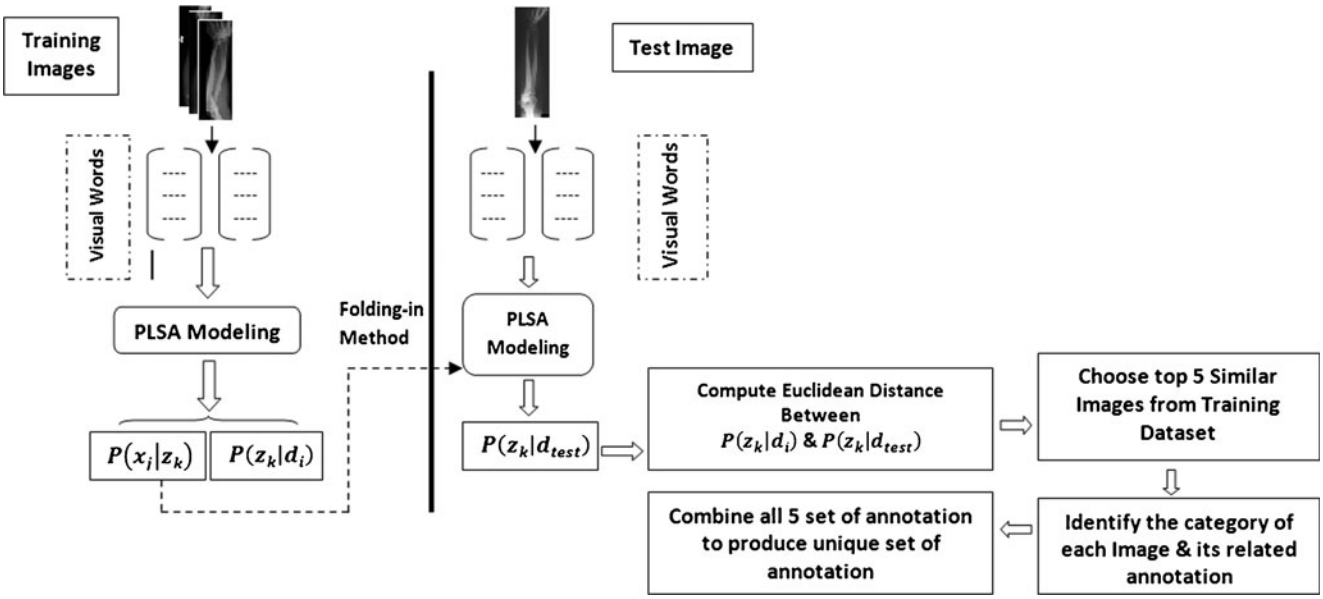

Fig. 4 The flow of retrieving similar images using PLSA

Two selection criteria were conducted on these keywords to generate the final annotation. Those first level keywords which appear in all the three sets generated from the above techniques were selected to fulfill the first criteria. Those keywords from the second level which are generated in any two sets were taken to perform the second selection criteria. The combination of all the selected keywords is the final annotation for the query image.

Figure 5 is the sample screenshot that represents this process. The three sets of produced keywords are divided into two levels according to their importance. Each keyword was given a unique number as shown in Fig. 2. These numbers are used to determine if the annotated words belong to level one or level two. Next, the keywords are loaded into the respective list box as shown in Fig. 5. The selection criteria are applied upon clicking on "Compute Final Annotation." As can be seen from the screenshot, the first level keywords "arm, forearm, wrist joint" are common in all the three sets and elbow joint, radius, ulna, left and AP view that appeared in any two sets of the keywords were taken as final annotation.

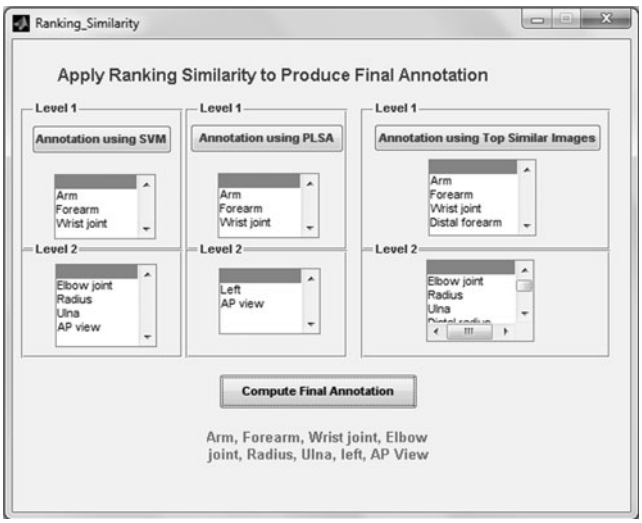

Fig. 5 Interface to represent the process of producing final annotation

Classification

This is the final phase of the proposed classification framework. In this phase, the test images are classified into respected classes through their annotation keywords generated from the previous section. This is done by computing the *Total Weight* (TW) of the generated annotation based on the following equation:

$$\mathrm{TW} = BR(L_1X + L_2Y + L_3Z) \tag{4}$$

Firstly, the keywords from each body region are divided into three levels; the first level contains those keywords clearly representing the body region. Unlike the annotation phase, the keywords related to imaging direction and view are separated from the second level and formed as level three. *X, Y*, and *Z* represent the three different levels of keywords. A different weightage is given to every level of keywords. The weights given to level one and level

Table 1 Weight assigned to each body region (BR)

	Abdomen	Arm	Leg	Chest	Cranium	Spine
Weight	10	20	30	40	50	60

Table 2 Weightage calculation and classification process from selected body region

Ground Truth	Arm, distal forearm, distal radius, distal ulna, left, lateral view	Cranium, facial cranium, orbits, skull, AP view
Level 1	Arm, distal forearm	Cranium, facial cranium
Weight	$L_1X=2\times3$	$L_1X=2\times3$
Level 2	Distal radius, distal ulna	Orbits, skull
Weight	$L_2Y=2\times2$	$L_2Y=2\times2$
Level 3	Left, lateral view	AP view
Weight	$Z_{\text{left}}=0.25$ $Z_{\text{lateral view}}=0.21$	$Z_{\text{AP View}}=0.33$
Weight of annotation without body region ($L_1X+L_2Y+L_3Z$)	6+4+(0.25+0.21)=10.46	6+4+0.33=10.33
Total weight including weight of body region:$BR(L_1X+L_2Y+L_3Z)$	20(10.46)=209.2	50(10.33)=516.5

two are $X=3$ and $Y=2$, respectively. Variable Z represents the third level keywords which are imaging view and direction. Examples of such keywords are lateral view, coronal view, left, right, etc. These keywords are not specific to any body region, and some of them may be common in most of the images from different body regions; therefore Z, is calculated as follows:

$$Z_{i1\ldots i9} = \frac{\text{Total number of keyword}_{i1\ldots i9} \text{ occured in training set}}{\text{Total number of training images}} \quad (5)$$

Variables L_1, L_2, and L_3 are the number of keywords from level one, level two, and level three appeared in its annotation, respectively.

The weight given for each level of the keywords is common for all the body regions. As demonstrated in Table 2, the weightage computed for given ground truth annotation for two different body regions is almost similar without multiplying with BR (the weight assigned for each body region) value. As such, in order to distinguish the body region from one another, a different value is allocated for each body region as shown in Table 1. The resulting value would help to differentiate the body region from one another more clearly. The weight for body region is represented by "BR" in Eq. 4.

Thus, the TW_{Test} calculated for each test image will be compared with the TW from each category of the training set and then the test image classified to the category with closest weightage to TW_{Test}.

Table 2 shows the calculation of TW and classification process of sample X-ray image from selected body regions. Every image category carries a different weight calculated based on its ground truth annotation. The same approach is used to compute the weight of the annotation made for the unseen test image. The computed weight of the test image is then compared with all the weight obtained from training dataset, and then the test image is classified to the category with closest weight to the test image's weight.

Experimental Result

In this section, a set of experiments were conducted to evaluate the performance of the classification algorithm on ImageCLEF 2007 medical dataset. This experiment specifically conducted on those classes left with lower accuracy rate in previous experiment [30]. These classes were labeled as low accuracy classes (LAC) and contain 39 classes.

Parameter Optimization

LIBSVM software package has been utilized to perform discriminative-based classification with nonlinear RBF kernel functions. The optimum kernel γ and cost C parameters have identified empirically with fivefold cross-validation. We use one-vs-one multi-class extension for Support Vector Machine

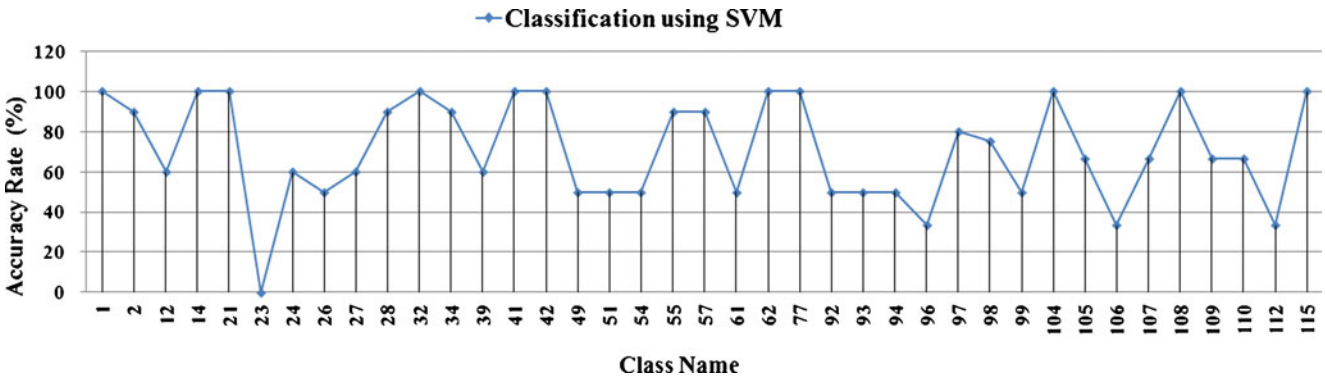

Fig. 6 Accuracy rate obtained on 39 classes using SVM

Table 3 Two levels of keywords belong to "Arm" body region

Level 1	Arm, wrist joint, shoulder joint, distal forearm, forearm, and carpal bone
Level 2	Scaphoid, scapula, distal radius, distal ulna, humerus, elbow joint, radius, ulna, upper humerus, Left, right, oblique view, lateral view, PA view, AP view, and axial view

(SVM). The main parameter in construction of bag of visual words is the number of visual words V in the visual vocabulary. Different vocabulary size has been considered starting from 100 followed by 200, 300, 400, 500, 600, and 700 to investigate how the classification performance is affected. The best performance obtained from $V=500$ in this study. Another parameter used in this experiment is the number of keywords used to construct the term–document matrix for textual vocabulary. These keywords are from LAC which consists of 39 classes. Totally, 68 unique keywords were identified after eliminating the duplicate keywords. The measurement used to evaluate classification performance is average accuracy or also referred as accuracy rate. The measurements used to evaluate the performance of the annotation are recall and precision. Therefore, annotation recall and precision are computed for every word in the testing set. Recall and precision are averaged over the set of testing words. In the case of automatic image annotation, the aim is to get both high recall and precision.

$$\text{Recall} = \frac{\text{number of images annotated correctly with a given word}}{\text{number of images that have that word}} \quad (6)$$

$$\text{Precision} = \frac{\text{number of images annotated correctly with a given word}}{\text{number of images annotated with that particular word}} \quad (7)$$

Classification Results on Low Accuracy Classes

In this section, we analyze and evaluate the classification performance on the unseen test image. The first section of the proposed classification algorithm is annotation. There were three different techniques used to perform annotation. The first technique was based on the supervised classification. As such, we follow the classification model constructed on LAC in [30]. That model classifies the unseen test image into one of the pre-defined classes. Figure 6 shows the classification results obtained by the model generated from these classes. The average accuracy rate reported was 72 %.

Then, the ground truth annotated keywords of every class are assigned to the test image accordingly to produce the first set of annotation. The average recall and average precision of the annotation made by this approach are 0.79 and 0.80, respectively.

For the second annotation techniques, linked pair of PLSA model was applied on the extracted BoW of the test images. As explained in "Annotation Using PLSA," the top five words were selected as annotation keywords for the unseen test images. The average recall and average precision of the annotation made by this approach are 0.77 and 0.78, respectively.

For the third annotation technique, PLSA model was applied on the extracted BoW of the unseen test images. Next, the top five similar images to the test images were selected from the training dataset. The respective class labels for these five images are known because they are from the training dataset. As such, the related keywords of each one of those classes are assigned to the test images. They are then combined to generate the unique annotation for each unseen test image. The average recall and average precision of the annotation made by this approach is 0.85 and 0.86, respectively.

Subsequently, ranking similarity need to be applied on each set of annotation produced by the above three techniques to construct the final annotation. To do this, annotated keywords are divided into two levels based on their importance. Level one consists of those keywords that clearly represent the body region, and level two contains those keywords that describe specific bone structure in the body region. Table 3 shows two levels of the keywords belonging to category of "arm." The average recall and average precision of the final annotation made after applying ranking similarity is 0.93 and 0.94, respectively.

Upon construction of final annotation, the Total Weightage (TW_{Test}) of the annotated keywords is calculated for every

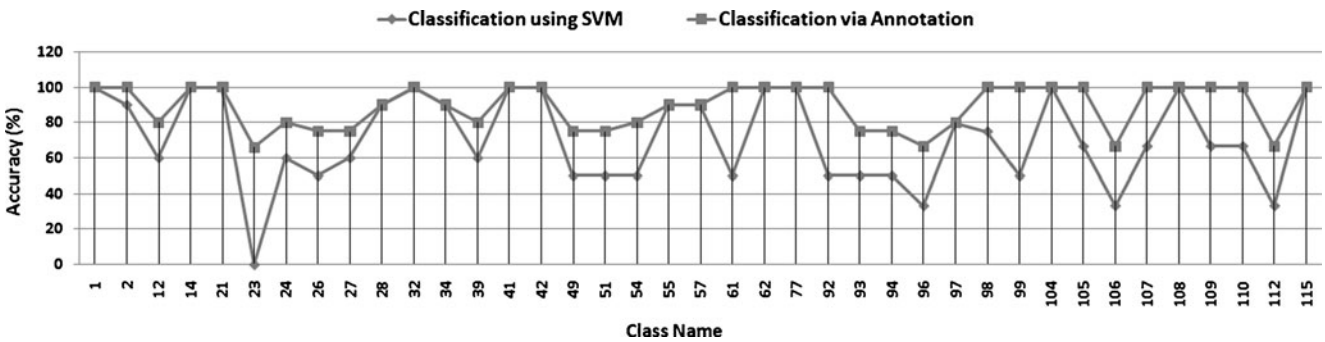

Fig. 7 Comparison on classification result on LAC obtained by SVM and proposed annotation

Table 4 Annotation and classification results on selected test image from class 23

Query Test Image			
Ground truth	Arm, forearm, wrist joint, elbow joint, radius, ulna, left, and AP view	Arm, forearm, wrist joint, elbow joint, radius, ulna, left, and AP view	Arm, forearm, wrist joint, elbow joint, radius, ulna, left, and AP view
Weight	311.60	311.60	311.60
Class no.	23	23	23
Annotation using Binary classification (SVM)	*Arm, forearm, wrist joint, elbow joint, radius, ulna, and AP view*	*Arm, forearm, wrist joint, elbow joint, radius, ulna, and AP view*	*Arm, forearm, wrist joint, elbow joint, radius, ulna, and AP view*
Annotation using PLSA	*Arm, forearm, wrist joint, distal forearm, distal radius, distal ulna, and left*	*Arm, forearm, wrist joint, left, and AP view*	*Arm, forearm, wrist Joint, right, and AP view*
Top similar image 1	Arm, forearm, wrist joint, elbow joint, radius, ulna, left, and AP view	Arm, forearm, wrist joint, elbow joint, radius, ulna, left, and AP view	Arm, forearm, wrist joint, elbow joint, radius, ulna, left, and AP view
Top similar image 2	Arm, forearm, wrist joint, elbow joint, radius, ulna, and AP view	Arm, distal forearm, distal radius, distal ulna, Left, and AP view	Arm, Distal forearm, Distal Radius, Distal Ulna, Left, and AP View
Top similar image 3	Arm, forearm, wrist joint, elbow joint, radius, ulna, and AP view	Arm, forearm, wrist joint, elbow joint, radius, ulna, and lateral view	Arm, forearm, wrist joint, elbow joint, radius, ulna, and lateral view
Top similar image 4	Arm, forearm, wrist joint, elbow joint, radius, ulna, and lateral view	Arm, forearm, wrist joint, elbow joint, radius, ulna, and AP view	Arm, forearm, wrist joint, elbow joint, radius, ulna, and AP view
Top similar image 5	Arm, forearm, wrist joint, elbow joint, radius, ulna, right, lateral view	Arm, distal forearm, distal radius, distal ulna, left, and AP view	Arm, distal forearm, distal radius, distal ulna, left, and AP View
Annotation using top similar images	*Arm, forearm, wrist joint, elbow joint, radius, ulna, left, right, lateral view, and AP view*	*Arm, forearm, wrist joint, elbow joint, radius, ulna, distal forearm, distal radius, distal ulna, left, lateral view, and AP view*	*Arm, forearm, wrist joint, elbow joint, radius, ulna, distal forearm, distal radius, distal ulna, left, lateral view, and AP view*
Final annotation	Arm, forearm, wrist joint, elbow joint, radius, ulna, left, and AP view	Arm, forearm, wrist joint, elbow joint, radius, ulna, left, and AP view	Arm, forearm, wrist joint, elbow joint, radius, ulna, and AP View
Weight	311.60	311.60	306.6
Class no.	23	23	96

The sample results of 3 annotation techniques are presented, as such they are typed with italic font

unseen test images using Eq. (4). Thus, the computed TW_{Test} will be compared with the TW from each category of the training set, and then the test image classified to the category with closest weightage to TW_{Test}. The average accuracy rate reported by this approach on 39 classes under LAC is 87.5 %. In Fig. 7, the classification performance obtained by this approach is compared with the result obtained by supervised SVM model for every individual class. The results show an improvement in classification performance as compared with single SVM classifier.

The annotation and classification results on the unseen test images from class 23 are further illustrated in Table 4. This class contains three test images. As shown in Fig. 7, we got two images classified correctly with the proposed method whereas the zero accuracy rates were reported using SVM classifier.

Discussion

The aim of this study is to improve the classification performance on LAC classes obtained in previous work [30].

In this experiment, we gave a special attention to those classes with high ratio of intraclass variability and interclass similarities. In order to explain how the proposed annotation framework rectifies the abovementioned problems, drill down analysis has been applied on those classes with high ratio of intraclass variability and interclass similarities. After the first iteration in previous work [30], 39 classes were left with accuracy below 80 %. Figure 6 shows the classification results obtained by the model generated from these classes. The average accuracy rate reported was 72 %.

Even though the number of classes involved in the model generated from LAC is lesser, it obtains a good classification performance. As can be seen in Fig. 6, there are 13 classes with accuracy below 60 %.

We have done detailed investigation on these classes to know the reason of low accuracy rate. Seven of them belong to "arm" body region. In ImageCLEF 2007 dataset, there are 33 classes under "Arm" body region. These classes are distributed into six sub-body regions. Based on the result obtained from Fig. 6, the seven classes with accuracy below 60 % belong to three sub-body region of "arm" as shown in Table 5.

In Table 6, the confusion matrix is created for these three sub-body regions. As can be seen, out of 12 test images in category of "forearm," 10 of them were classified correctly.

Table 5 Number of classes in every sub-body regions of "Arm"

Sub-body region	Number of classes
Forearm	4
Shoulder	1
Hand	2

Table 6 Confusion matrix on sub-body region of "Arm"

	Forearm	Shoulder	Hand	Other region	Accuracy rate (%)
Forearm	10		1	1	83
Shoulder		4		1	80
Hand			8	1	89

One of the test images from the categories of hand and shoulder was misclassified as presented in confusion matrix.

The number of classes for each sub-body region of "Arm" is listed in Table 7. We also show the number of classes of this sub-body region with accuracy rate of 60 % and above in Table 7. As can be seen from Table 7, none of the four classes under "forearm" sub-body region could managed to attain accuracy rate of 60 % even though the accuracy rate of this sub-body region was reported 83 % in Table 6. This analysis represents the high ratio of misclassification among classes under "forearm" body region. This misclassification can be seen in the other two sub-body region of "Arm."

High ratio of interclass similarities and intraclass variability among these classes is the main reason of misclassification. Inspired from this fact, we proposed a classification framework which utilizes the annotated keywords of the images to improve the classification performance.

Annotation module in the proposed framework plays a very important role in the proposed classification algorithm; a good performance in annotation would improve the classification performance. In Table 4, annotation and classification results on test images from class 23 are represented. There are three test images provided for class 23 that are misclassified as class 96 using SVM. In both classes 23 and 96 referring to "forearm" sub-body region, they are distinguished from each other only in direction. Class 23 has the direction "left" in its annotation, but no direction stated for class 96. Meaning that in the case of annotation using classification, the corresponding annotated keywords from both classes are almost similar.

As for annotation using top similar images, mostly the top five retrieved similar images to the query image are belonging to the same class or same sub-body region. Thus, they are sharing most of the important keywords. In the case of annotation using

Table 7 Number of classes per each sub-body regions

	Number of classes	Number of classes with accuracy of above 60 %
Forearm	4	0
Shoulder	1	0
Hand	2	0

PLSA, a linked pair of PLSA models is employed to capture semantic information from textual and visual modalities and learn the correlation between them. It is clear that this structure can predict most of the important keywords (level one) correctly. Therefore, the combined set of keywords generated from this annotation technique would contain most of the keywords of the respective sub-body region.

The experimental results obtained on the entire database shows an improvement in probability of getting more accurate annotation by fusing the above three techniques which would lead to an increment in classification accuracy. By observing the results obtained from Table 4, it is clearly evident that all the three techniques in particular, annotation using PLSA and annotation using top similar images, certain classes can be effectively used to annotate correctly and accurately compared to SVM classification because it incorporate both textual and visual features of the images. Accuracy rate obtained by the proposed annotation algorithm shows tremendous improvement compared to classification rate obtained by SVM as illustrated in Fig. 7.

Conclusion

In this paper, a classification framework is proposed to improve the accuracy rate of those classes of medical X-ray images with great intraclass variability and interclass similarities. This classification task carried out by employing three different annotation techniques such as annotation by binary classification, PLSA-based image annotation, and annotation using top similar images to the query image. The final annotation is then constructed by utilizing ranking similarity on annotated keywords. Next, the final annotation is used for classification purpose by computing their weightage and comparing with each category's weightage in database. The experimental result shows that the accuracy rate obtained outperformed those works stated in the literature review.

Acknowledgments The authors would like to thank Thomas Deserno, Department of Medical Informatics, RWTH Aachen, Germany, for making the database available for the experiments.

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Güld MO, Kohnen M, Keysers D, Schubert H, Wein B, Bredno J, Lehmann TM: Quality of dicom header information for image categorization. Proc SPIE Int Symp Med Imaging 4685:280–287, 2002
2. Rui Y, Huang TS: Image Retrieval: Current Techniques, Promising, Directions, and Open Issues. J Vis Commun Image Represent 10:39–62, 1999
3. Müller H, Deselaers T, Kim E, Kalpathy-Cramer J, Deserno TM, Clough P, Hersh W: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Working Notes of the 2007 CLEF Workshop, 2007
4. Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB: The IRMA code for unique classification of medical images. In: Proceedings SPIE, 5033: 440–451, 2003
5. Bo Q, Wei XIONG, Qi TIAN, Xu CS: Report for Annotation task in ImageCLEFmed 2005. In: working notes of CLEF 2005. (Vienna, Austria), 2005
6. Muller H, Geissbuhler A, Marty J, Lovis C, and Ruch P: The use of MedGIFT and EasyIR for ImageCLEF 2005. In: CLEF 2005 Proceedings. Lecture notes in computer science (LNCS), 4022: 724–32, 2005
7. GuangJian T, Hong F, Feng DD: Automatic medical image categorization and annotation using LBP and MPEG-7 edge histograms. Int. Conf. Information Technology and Applications in Biomedicine, 51–53, 2008
8. Mueen A, Zainuddin R, Sapiyan Baba M: Automatic Multilevel Medical Image Annotation and retrieval. J Digit Imaging 21(3): 290–295, 2008
9. Tommasi T, Orabona F, Caputo B: Discriminative cue integration for medical image annotation. Pattern Recogn Lett 29:1996–2002, 2008
10. Zhy C-M, Gu G-C, Liu H-B, Shen J, Yu H: Segmentation of Ultrasound Image Based on Texture Feature and Graph Cut. Int Conf Comput Sci Softw Eng 1:795–798, 2008
11. Jeanne V, Unay D, Jacquet V: Automatic Detection of body parts in X-ray images. IEEE Comput Soc Conf Comput Vis Pattern Recognit, 25–30, 2009
12. Pourghasem H, Ghasemian H: Content-based medical image classification using a new hierarchical merging scheme. Comput Med Imaging Graph 32(8):651–661, 2008
13. Rahman MM, Desai BC, Bhattacharya P: Medical Image Retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion. Comput Med Imaging Graph 32(2):95–108, 2007
14. Dimitrovski I, Kocev D, Loskovska S, Džeroski S: Hierarchical annotation of medical images. Pattern Recogn 44:2436–2449, 2011
15. Ko B, Kim S, Nam J-Y: X-ray Image Classification Using Random Forests with Local Wavelet-Based CS-Local Binary Patterns. J Digit Imaging 24:1141–1151, 2011
16. Rahman MM, Antani SK, Thoma GR: A query expansion framework in image retrieval domain based on local and global analysis. Inf Process Manag 47:676–691, 2011
17. Unay D, Soldea O, Ekin A, Cetin M, Ercil A: Automatic annotation of X-Ray images: a study on Attribute Selection, In Medical Content-Based Retrieval for Clinical Decision Support, 5853: 97–109, Springer Berlin Heidelberg, 2010
18. Haralick RM, Shanmugam K, Dinstein I: Textural features for image classification. IEEE Trans Syst Man Cybern 3(6):610–621, 1973
19. Ojala T, Pietikainen M, Harwood D: A Comparative Study of Texture Measures with Classification Based on feature Distributions. Pattern Recogn 29(1):51–59, 1996
20. Lazebnik S, Schmid C, Ponce J: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. IEEE Comput Soc Conf Comput Vis Pattern Recognit 2:2169–2178, 2006
21. Jie F, Jiao LC, Xiangrong Z, Dongdong Y: Bag-of-Visual-Words Based on Clonal Selection Algorithm for SAR Image Classification. IEEE Geosci Remote Sens Lett 8:691–695, 2011
22. Teng L, Tao M, In-So K, Xian-Sheng H: Contextual Bag-of-Words for Visual Categorization. IEEE Trans Circuits Syst Video Technol 21:381–392, 2011
23. Sui L, Zhang J, Zhuo L, Yang YC: Research on pornographic images recognition method based on visual words in a compressed domain. IET Image Process 6:87–93, 2012

24. Yang F, Lu H, Zhang W, Yang G: Visual tracking via bag of features. IET Image Process 6:115–128, 2012
25. Deselaers T, Hegerath A, Keysers D, Ney H: Sparse Patch-Histograms for Object Classification in Cluttered Images. Pattern Recogn, 4174: 202–211. Springer Berlin Heidelberg, 2006
26. Avni U, Jacob G, Michal S, Eli K, Hayit G: Chest x-ray characterization: from organ identification to pathology categorization. In: Proceedings of the international conference on Multimedia information retrieval. ACM, Philadelphia, 2010, pp 155–164
27. Avni U, Greenspan H, Konen E, Sharon M, Goldberger J: X-ray Categorization and Retrieval on the Organ and Pathology Level, Using Patch-Based Visual Words. IEEE Trans Med Imaging 30: 733–746, 2011
28. Wei Y, Zhentai L, Mei Y, Meiyan H, Qianjin F, Wufan C: Content-Based Retrieval of Focal Liver Lesions Using Bag of Visual Words Representations of Single and Multiphase Contrast Enhanced CT Images. J Digit Imaging 25:708–719, 2012
29. Deselaers T, Ney H: Deformations, patches, and discriminative models for automatic annotation of medical radiographs. Pattern Recogn Lett 29:2003–2010, 2008
30. Zare MR, Mueen A, Woo CS: Automatic Classification of Medical X-ray Images using Bag of Visual Word. IET Comput Vis 7(2):105–114, 2013
31. Zare MR, Woo CS, Mueen A: Automatic Classification of medical X-ray Images. Malays J Comput Sci 26(1):9–22, 2013
32. Monay F, Gatica-Perez D: Modeling semantic aspects for cross-media image indexing. IEEE Trans Pattern Anal Mach Intell 29(10):1802–1817, 2007
33. Li Z, Shi Z, Liu X, Shi Z: Modeling continuous visual features for semantic image annotation and retrieval. Pattern Recogn Lett 32(3): 516–523, 2011
34. Chandrika P, Jawahar CV: Multi modal semantic indexing for image retrieval. Proc. of the ACM Int'l Conf. on Image and Video Retrieval, 342–349, 2010
35. Hoffman T: Unsupervised learning by probabilistic latent semantic analysis. J Mach Learn 42(1–2):177–196, 2001
36. Quelhas P, Monay F, Odobez JM, Gatica-Perez T, Tuytelaars T, Van Gool L: Modeling scenes with local descriptors and latent aspects. IEEE Int Conf Comput Vis, 883–890, 2005
37. Zare MR, Mueen A, Awedh M, Woo CS: Automatic classification of medical X-ray images: hybrid generativediscriminative approach. IET Image Process 7(5):523–532, 2013
38. Lowe D: Distinctive image features from scale invariant key points. Int J Comput Vis 60(2):91–110, 2004
39. Lowe D: Object recognition from local scale-invariant features. Int Conf Comput Vis 2:1150–1157, 1999