

Automatic medical image annotation and retrieval

Jian Yao^{a,*}, Zhongfei (Mark) Zhang^a, Sameer Antani^b, Rodney Long^b, George Thoma^b

^aDepartment of Computer Science, State University of New York at Binghamton, Binghamton, NY 13902, USA

^bNational Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Available online 4 March 2008

Abstract

The demand for automatically annotating and retrieving medical images is growing faster than ever. In this paper, we present a novel medical image retrieval method for a special medical image retrieval problem where the images in the retrieval database can be annotated into one of the pre-defined labels. Even more, a user may query the database with an image that is close to but not exactly what he/she expects. The retrieval consists of the *deducible retrieval* and the *traditional retrieval*. The deducible retrieval is a special semantic retrieval and is to retrieve the label that a user expects while the traditional retrieval is to retrieve the images in the database which belong to this label and are most similar to the query image in appearance. The deducible retrieval is achieved using SEMI-supervised Semantic Error-Correcting output Codes (SEMI-SECC). The active learning method is also exploited to further reduce the number of the required ground truthed training images. Relevance feedbacks (RFs) are used in both retrieval steps: in the deducible retrieval, RF acts as a short-term memory feedback and helps identify the label that a user expects; in the traditional retrieval, RF acts as a long-term memory feedback and helps ground truth the unlabelled training images in the database. The experimental results on IMAGECLEF 2005 [<http://ir.shef.ac.uk/imageclef2005/>] annotation data set clearly show the strength and the promise of the presented methods.

© 2008 Elsevier B.V. All rights reserved.

Keywords: Medical image; Image retrieval; Image annotation; Semi-supervised learning; Active learning; Error-correcting output coding

1. Introduction

Medical images play a central role in patient diagnosis, therapy, surgical planning, medical reference, and training. With the advent of digital imaging modalities, as well as images digitized from conventional devices, collections of medical images are increasingly being held in digital form. It becomes increasingly expensive to manually annotate medical images. Consequently, automatic medical image annotation becomes important.

Due to the large number of the images without text information, content-based medical image retrieval (CBMIR) has received increased attention. We call the semantic similarity defined between different appearances of the same object the *intra-object similarity* and the semantic similarity defined between different objects the *inter-object similarity*. A semantic similarity in this paper

refers to both intra-object and inter-object semantic similarities. Each image in the database contains only one object. The semantic similarity between two images is the semantic similarity between the objects contained by the images. For example, the semantic similarity between an elbow image in coronal view and an elbow image in sagittal view is the intra-object similarity while the semantic similarity between a hand image and an upper-arm image is the inter-object similarity.

It is well known that CBMIR is quite different from CBIR as the retrieval similarity must consider the medical context (such as the subtle pathological changes) as well as the user individualized subjectivity. On the other hand, the medical context in CBMIR is often addressed through automatic medical image annotation, which is a special scenario of the general image annotation problem as the annotation vocabulary consists of all the expected image labels in a specific application. The problem addressed in this paper is a special medical image retrieval problem. Compared with the general medical image retrieval

*Corresponding author.

E-mail address: jian_yao1414@yahoo.com (J. Yao).

problems, the problem addressed here has the following properties:

1. The images in the retrieval database can be annotated into one of the pre-defined labels, which are denoted as the *ground truth labels* of the images. Due to the ground truthing complexity, only a small portion of the whole image collections have their ground truth labels available.
2. Given a specific query, the correctly retrieved images should have the same ground truth label, which may not necessarily equal to the ground truth label of the query image provided that the query image and the retrieved images share a sufficient semantic similarity. This means that a user may query the database with an image that is close to but not exactly what he/she expects.

2. Related work

Medical image annotation is a special classification, i.e., classifying a given image into one of the pre-defined labels. Consequently, existing classification methods [3,6–8,11,13,14,27] may be applied to the medical image annotation. Due to the limitation of the availability of ground truthed training samples, semi-supervised learning (SSL) methods [17,21,28,29] have received significant attention recently. It only requires a small ground truthed training set, together with a large unlabelled training set. Typical SSL methods ground truth training samples randomly while semi-supervised active learning methods [12,20] selectively ground truth training samples. It is reported [20] that semi-supervised active learning methods usually need less ground truthed training samples than SSL methods.

Annotation typically has a large number of possible labels. For example, the number of the labels for the data set from IMAGECLEF 2005 annotation task is 57. As is well known in machine learning community that when the number of labels becomes large, not only training a classifier to directly solve the classification problem is expensive, but also the learned classifier tends to have a poor performance. Consequently, indirect classification methods such as Error-Correcting Output Codes (ECOC) [5,9] may be used. ECOC solves a multi-class classification problem by solving a set of two-class classification problems. Unfortunately, existing ECOC methods are not semantically based, leading to the difficulty to exploit the semantic similarity between different images. Furthermore, to the best of our knowledge, there is no SSL version or active learning version ECOC method proposed in literature.

Existing medical image retrieval methods [4,10,15,18,23] are to retrieve the images in a database which are most similar to a query image. Most methods focus on the appearance-based similarity, i.e., the appearance of the retrieved images is similar to that of the query image. There

is little semantic information exploited. Among the few efforts that claim to have the semantic information exploited, the semantic similarity is defined between different appearances of the same object, i.e., the intra-object similarity. Consequently, it would be considered as an incorrect retrieval for the existing CBMIR methods to retrieve coronal foot images given an axial knee query image.

Relevance feedback (RF) [16,19,22,24–26] has been an active research area in CBMIR and CBIR. RF intends to bridge the gap between the low-level image features and the high-level image semantics by analyzing and employing the feedback information. To the best of our knowledge, in the literature RF is not yet used to bridge the gap of the semantic differences, especially the inter-object semantic difference, between the query image and the expected retrieval images.

The main contributions of this paper are listed as follows:

1. We define the concept of the semantic similarity between different images and develop a retrieval method under this semantic similarity. Users may pose query images that are close to but not exactly what they expect.
2. We propose a novel SEMI-SECC annotation method, which is a semantic ECOC under SSL and active learning frameworks.
3. RFs are used in the retrieval method which not only help identify what a user expects but also help discover the ground truth for unlabelled training samples.

3. Annotation model

In this section, we first give a brief introduction to the ECOC method; we then revise ECOC to develop a semantic ECOC–SECC. SECC model consists of two steps: the individual classifications and the combination of the results from individual classifications. Finally, we present the SSL version SECC–SEMI-SECC, which also exploits active learning method to further reduce the number of the required ground truthed training samples.

3.1. Error-correcting output codes (ECOC)

ECOC is used to solve a multi-class classification problem using multiple two-class classifiers, which are called *individual classifiers*. The procedure to select the individual classifiers is called *coding*. The labels of the original multi-class classification problem are called *overall labels*. The labels of the individual classifiers are called *individual labels*. If we represent the individual labels of one sample as a vector, which is called the *code* of the sample, all the training samples with the same overall label should have the same code. Table 1 gives a simple example, where there are four overall labels: forearm and sagittal, elbow and coronal, foot and axial, and foot and sagittal. Four individual classifiers are used in an ECOC solution.

Table 1
A simple classification problem together with its ECOC and SECC coding

Overall label ID	ECOC codes	SECC codes
0 (forearm and sagittal)	(1,0,1,0)	(1,0,1)
1 (elbow and coronal)	(1,1,1,1)	(2,0,2)
2 (foot and axial)	(0,1,0,0)	(0,1,0)
3 (foot and sagittal)	(0,0,1,1)	(0,1,1)

The criterion of ECOC coding is that the difference between the codes of different overall labels should be sufficiently large, which is typically measured using the Hamming distance. Typically, the individual classifiers are randomly selected and the more individual classifiers, the higher accuracy the overall classifier has. ECOC classification is solved by finding the code whose distance to the query code is the minimum. In the above example, if a query has a code (1, 1, 0, 0), it will be classified to “Label ID 2” since the corresponding Hamming distance is smaller than those of the query code to the other codes. In the following text, we explain how the proposed method selects the individual classifiers and finds the closest code, i.e., combines the individual classifiers.

3.2. SECC individual classifiers’ selection (coding)

A typical overall label for IMAGECLEF 2005 annotation data set is “elbow image, sagittal view, plain radiography, and musculoskeletal”. We denote each part of an overall label as a *category* and the possible values for that category as *category labels*. For the example given in Table 1, we may define three categories: Arm (possible labels: forearm, elbow, and non-arm), Foot (possible labels: foot and non-foot), and View (possible labels: axial, sagittal, and coronal). In some applications, not only the overall label related information but also the category related information are required to be determined. Since the individual classifiers in ECOC coding are selected randomly, they seldom contain the latter information. Regarding the ECOC solution given in Table 1, it is unlikely that an individual classifier would solve the classification problem w.r.t. one of the three categories exactly. In order to determine the category related information, we propose to revise ECOC to SECC as follows.

First, we define several categories and category labels for a data set. Categories independent of other categories are called *independent categories*. In the above example, the View category is in general independent of other categories. Categories correlated to other categories are called *correlated categories*. The Arm category and the Foot category in the above example are correlated. An image with a forearm category label can only have a non-foot category label. Each correlated category has several labels corresponding to different aspects of the category, together with a “non-” label. A sample with a “non-” label in a

category means that the sample does not belong to that category. In the above example, if a sample has a “non-arm” label, this sample is not part of an arm. The label ID for a “non-” label is 0 while those for the remaining category labels are non-zero values. Note that for one sample, there is only one correlated category such that the category label of the sample on this category is not a “non-” label. This category is called the *delegate* category of the sample.

We then train one individual classifier for one category. This classifier may be a two-class classifier; it may also be a multi-class classifier. Different individual classifiers may use different classification models and different feature sets. Table 1 also gives a possible SECC coding solution. Since each individual classifier focuses on one category in SECC, we do not distinguish between the individual label and the category label in the following text.

3.3. SECC individual classifiers’ combination

It is clear that SECC coding does not guarantee that the difference between the codes of different overall labels is sufficiently large. Consequently, the ECOC similarity functions (e.g., the Hamming distance function) may not be suitable for SECC. Here we present a probabilistically based similarity function for SECC. Let the number of the individual classifiers be M . Let the number of the different individual labels for individual classifier j be M_j . Let a query image be x_i . Denote the probability for x_i to have individual label k on individual classifier j as q_i^{jk} . Let $Q_i = \{q_i^{jk}\}$. Denote a possible code for x_i as $Y = (y^1, y^2, \dots, y^M)$ and the code of overall label o as $G_o = (g_o^1, g_o^2, \dots, g_o^M)$. We maximize the joint probability of G_o and Y given Q_i to find the overall label of the query image:

$$\text{Max}_{o, Y} P(G_o, Y|Q_i) = P(G_o|Y, Q_i) \times P(Y|Q_i), \quad (1)$$

where $P(Y|Q_i)$ is the probability of the event that the individual classification results are y^j 's given Q_i . Different individual classifiers are trained independently. Thus, it is possible that for some Y , the number of the non-zero y^j 's for correlated categories is not 1. Note that this is in conflict with the requirement that there is only one delegate category. Consequently, the corresponding $P(Y|Q_i)$ is set to 0. For other situations, $P(Y|Q_i)$ is set to the multiplication of the probabilities that the individual classification labels are correct, i.e., $q_i^{y^j}$'s. Let y^{C_j} 's be the y^j 's for the correlated categories. We then define $P(Y|Q_i)$ as follows:

$$P(Y|Q_i) = \begin{cases} 0, & |\{y^{C_j}, y^{C_j} \neq 0\}| \neq 1, \\ \prod_{j=0}^{M-1} q_i^{y^j}, & |\{y^{C_j}, y^{C_j} \neq 0\}| = 1. \end{cases} \quad (2)$$

$P(G_o|Y, Q_i)$ in Eq. (1) is the probability of the event that a query code Y with the probability set Q_i happens to be the ground truth code G_o . To simplify the computation, we let $P(G_o|Y, Q_i) = P(G_o|Y)$.

Let $D_o = \{|j, g'_o \neq y^j|\}$, i.e., the number of the y^j 's which are not equal to the corresponding g'_o . We then define $P(G_o|Y)$ as follows:

$$P(G_o|Y) = \begin{cases} 0, & D_o \geq T_1, \\ P(\{(j, g'_o), g'_o \neq y^j\} | \{(j, g'_o), g'_o = y^j\}), & D_o < T_1. \end{cases} \quad (3)$$

The conditional probability in the right-hand side of Eq. (3) is the probability of the event that when a query code contains part of the code of G_o , the remaining part of the query code happens to be the remaining part of the code of G_o . In order to focus the attention on the query codes that do not differ substantially from the code G_o , we introduce a threshold T_1 . If the code of G_o differs from the query code by at least T_1 bits, $P(G_o|Y)$ is set to 0. By assuming that each training image is identically and independently generated from an unknown distribution (i.i.d.), $P(\{(j, g'_o), g'_o \neq y^j\} | \{(j, g'_o), g'_o = y^j\})$ can be estimated using the training samples. For example, referring to the example in Table 1, assume that Label ID 0 has 20 training samples and Label ID 1 has 30 training samples. Since only Label ID 0 and Label ID 1 satisfy that $y^0 = 1$ and $y^2 = 1$, the probability of the event that $y^1 = 0$ and $y^3 = 0$ given the fact that $y^0 = 1$ and $y^2 = 1$ is determined as follows:

$$P(\{(1, 0), (3, 0)\} | \{(0, 1), (2, 1)\}) = \frac{20}{20 + 30}. \quad (4)$$

3.4. Semi-supervised active learning SECC

A typical SSL method works as follows: learn a supervised classifier using the ground truthed training samples only; label the unlabelled samples using the learned supervised classifier; re-train the supervised classifier using all the training samples. The last two steps are repeated until certain stop criteria are met. SEMI-SECC follows the enhanced semi-supervised learning (ESL) framework presented in [29]. The ESL framework is probabilistically guaranteed to have the accuracy increased when the number of iterations increases. The drawback of the ESL model, which is also true for general SSL methods, is that the ground truthed training sample selection is random. Consequently, it may require more ground truthed training samples than necessary to achieve an acceptable accuracy. Here we exploit active learning methods to economically ground truth the unlabelled training samples.

For unlabelled training sample x_i and individual classifier j , if the probabilities for x_i to belong to different individual labels on individual classifier j are close, it means that this sample is probably hard to classify by the current individual classifier j and thus should have a high priority to be ground truthed if individual classifier j requires more ground truthed samples. Consequently, we define the *uncertainty* for x_i

on individual classifier j using the entropy:

$$c_{ij} = - \sum_{k=0}^{M_j-1} q_i^{jk} \times \log(q_i^{jk}). \quad (5)$$

We use the accuracy, which is the percentage of the correctly annotated samples, to determine whether or not an individual classifier needs more ground truthed samples. When an individual classifier has a high accuracy, it means that it probably does not need more training samples, and vice versa. Let the estimated ground truth accuracy for individual classifier j be η^j , which can be estimated using either the test data or the method presented in [29]. We then define the *ground truthing priority* for x_i as

$$\gamma_i = \sum_{j=0}^{M-1} c_{ij} \times (1 - \eta^j). \quad (6)$$

In each iteration, we ground truth those unlabelled samples with the highest ground truthing priorities. The SEMI-SECC learning procedure is summarized in Algorithm 1. Since SEMI-SECC follows the ESL framework, which guarantees convergence [29], SEMI-SECC also converges.

Algorithm 1. SEMI-SECC learning procedure.

1. Ground truth a small set of images from the database.
2. Learn the initial individual classifiers. Set $i = 0$.
3. Set $i = i + 1$. Classify unlabelled samples using the trained classifiers at Iteration $i - 1$ and assign labels to unlabelled samples based on the classification results.
4. Determine γ_i for unlabelled sample x_i using Eq. (6) and ground truth the samples with the highest γ_i values.
5. Re-train the individual classifiers.
6. If certain stop criteria meet, stop. Otherwise, goto step 3.

4. Retrieval model

We present a two level retrieval model here. The first level is a special semantic retrieval, which is called the *deducible retrieval*. It is aimed to retrieve the overall label a user expects which is semantically similar to the overall label of the query image. The second level is a *traditional retrieval*, which is to retrieve images in a database with the overall label determined by the deducible retrieval. We first explain the deducible retrieval model and then introduce the traditional retrieval model.

4.1. Deducible retrieval

Since the deducible retrieval focuses on the semantic similarities, i.e., the semantic similarities for the same object or among different objects, we must specify such semantic similarities in advance. Unfortunately, such semantic similarities are subjective. For example, the same semantic similarity may be defined between different views of the same object, or between different parts of the same object, or between different objects. Thus, such semantic similarities should be dynamic instead of static.

The short-term memory RF is used to estimate such subjective semantic similarities. We only care about the specific semantic information that a user expects and do not care about the semantic information that the user does not expect. Therefore, only positive feedback is necessary in this case.

For each category j and category label k , we define a *user expected degree* α_{jk} , which is used to describe the degree a user expects on this category label. The summation of all the α_{jk} 's is 1. For the independent category labels and the correlated category labels that happen to be the “non-” category label, the initial user expected degrees are 0; for the correlated category labels different from the “non-” category label, the initial user expected degrees are non-zero. Let the code for query image x_i be Y and the code for the feedback image be G_o . Then, α_{jk} 's are updated as follows:

$$\alpha_{jk}^{\text{old}}, \quad y^j = g_o^j,$$

$$\alpha_{jk}^{\text{new}} = \begin{cases} 0, & y^j \neq g_o^j, \quad g_o^j = 0, \\ 0, & y^j \neq g_o^j, \quad g_o^j \neq 0, \quad k = 0, \\ 1, & y^j \neq g_o^j, \quad g_o^j \neq 0, \quad k = g_o^j, \\ 1/2, & y^j \neq g_o^j, \quad g_o^j \neq 0, \quad k \neq g_o^j, \quad k > 0. \end{cases} \quad (7)$$

The basic idea is that if there is no change for the label of a category, there is no change for the corresponding α_{jk} 's of this category. If a category label is changed to 0, the corresponding α_{jk} 's are all set to 0. Otherwise, the α_{jk} 's corresponding to this category are set to non-zero values with 1 for the α_{jk} 's corresponding to the changed category label and $\frac{1}{2}$ for the α_{jk} 's of other category labels except the “non-” category label. The α_{jk} 's corresponding to a “non-” category label are always 0. After this update, α_{jk} 's are linearly scaled so that the summation of them is 1. The semantic similarity between the query code Y and the ground truth code G_o is then defined as

$$R(G_o, Y) = \sum_{j=0, g_o^j \times y^j \neq 0}^{M-1} \alpha_{jg_o^j}. \quad (8)$$

It is not difficult to see that under the initial settings of α_{jk} 's, two overall labels are semantically similar to each other if their delegate categories are the same, i.e., the initial deducible retrieval results depend only on the classified delegate category of the query image. We then modify the optimization problem in Eq. (1) as

$$\text{Max}_{o, Y} P(G_o, Y|Q_i) \times R(Y, G_o). \quad (9)$$

For a query code Y , there are only a few G_o 's which have non-zero $R(G_o, Y)$ values. Consequently, there are only a few G_o 's which have non-zero similarities to the query code Y . Those G_o 's are the deducible retrieval results.

We want to mention that we assume that within a short period, the semantic information a user expects is constant. Consequently, the *user expected degree* sets within this short period remains same or at least same on the specific

category the user is interested. The deducible retrieval will then better return the label the user expects. For example, if a user expects arm images with coronal view while the system classify the query image as an arm image with sagittal view (the query image may or may not be an arm image with sagittal view), the *user expected degrees* corresponding to arm and coronal view will be increased after the short-term RF is done. As a result of fact, when the same user who expects foot images with coronal view queries the system by a foot image with sagittal view, the possibility that the user expects label, i.e., foot image with coronal view, is within the deducible retrieval results is increased.

4.2. Traditional retrieval

The traditional retrieval results are all from the overall label determined by the deducible retrieval, which is either the annotation overall label or the deducible retrieval feedback label. When there is no deducible retrieval feedback, the traditional retrieval results are the images in the database with the same annotation overall label as that of the query image and their appearance similarities to the query image are the largest. When there is a deducible retrieval feedback, the traditional retrieval results are the images randomly selected from the images in the database with the specified deducible retrieval feedback label.

Unlike the RF model for the deducible retrieval, the RF model for the traditional retrieval is aimed at recovering the ground truth of the unlabelled images in the database. Consequently, it is a long-term memory RF and it is necessary to have both positive and negative feedbacks. For each unlabelled training sample x_i , denote β_i^j as the probability that x_i belongs to overall label j . Assume that x_i is a positive traditional retrieval feedback image corresponding to overall label k , which is either the annotated ground truth label for x_i or the feedback label of the deducible retrieval. Typically, k is equal to the ground truth overall label of x_i . We update β_i^j 's as follows:

$$\beta_i^j = \begin{cases} 1, & j = k, \\ 0, & j \neq k. \end{cases} \quad (10)$$

If x_i is a negative traditional retrieval feedback image, we update β_i^j 's as follows:

$$\beta_i^j = \begin{cases} 0, & j = k, \\ \frac{\beta_i^j}{1 - \beta_i^k}, & j \neq k. \end{cases} \quad (11)$$

Due to the possibility that the ground truth overall label of x_i may not equal to the overall label a user expects, it is possible that k is not equal to the ground truth overall label of x_i . Consequently, the above updating procedure may not correctly update the ground truth. In order to avoid this scenario, if any of the following conditions happens, the ground truth update should not be applied: (1) k is not equal to any of the overall labels of the ground truthed

positive feedback images; (2) k is equal to any of the overall labels of the ground truthed negative feedback images; (3) the similarity between x_i and overall label k is less than a pre-defined threshold T_2 , which is empirically selected.

After the update, if for unlabelled sample x_i , there is only one non-zero β_i^j , we consider that overall label j is the ground truth overall label for x_i and add x_i to the ground truthed data set. For unlabelled training samples x_i with more than one non-zero β_i^j 's, Q_i is set based on β_i^j 's. After the number of conducted feedbacks reaches a threshold T_3 , the SEMI-SECC learning procedure is applied again using the updated ground truth and Q_i .

5. Experiments

5.1. Data set

The data set we use to evaluate our methods is the IMAGECLEF [1] 2005 annotation data set. All the images are X-ray images. There are 9000 training images and 1000 test images. These images can be categorized into 57 classes. Each class has 9–2563 training images. Fig. 1 displays 56 normalized images, with one image corresponding to one class. The images have different sizes before the normalization.



Fig. 1. Sample images from IMAGECLEF annotation database.

Table 2
Categories and category labels defined for IMAGECLEF 2005 annotation data set

Category	Category labels
Cranium (C)	Non-cranium, Cranium, Facial Cranium
Spine (C)	Non-spine, Cervical Spine, Thoracic Spine, Lumbar Spine
Arm (C)	Non-arm, Hand, Radio Carpal Joint, Forearm, Elbow, Upper Arm, Shoulder
Leg (C)	Non-leg, Foot, Ankle Joint, Lower Leg, Knee, Upper Leg, Hip
View (I)	Coronal, Sagittal, Axial, Others
Radiography (I)	Plain radiography, Fluoroscopy, Angiography
Function (I)	Musculoskeletal, Gastrointestinal, Uropoietic, Reproductive, Cardiovascular, Respiratory
Chest (C)	Non-chest, Chest, Chest Bone
Abdomen (C)	Non-abdomen, Abdomen, Upper abdomen
Pelvis (C)	Non-pelvis, Pelvis
Breast (C)	Non-breast, Left breast, Right breast

We define 11 categories for the data set. The categories and the category labels are listed in Table 2. The C and I in the category column represent the correlated category and the independent category, respectively.

5.2. Learning procedure

First, we normalize all the images to 16×16 size. Two hundred images are selected as ground truthed training images. Three different kinds of features—the intensity, the Harr wavelet feature, and the Garbor wavelet feature—are extracted from the normalized images. Algorithm 1 is then used to train the annotation model. Each individual classifier exploits only one of the three features. During each iteration of the learning, 20 unlabelled images with the highest γ_i values are selected and ground truthed. After the learning is finished, users are then asked to query the system using query images which may not match what they expect. The deducible retrieval feedback is provided when the annotation result does not match the user expected label. The user expected degrees are then updated as is discussed in Section 4.1. Users are then asked to provide several positive and negative feedbacks for the traditional retrieval. The probability for the unlabelled feedback images to belong to different overall labels are updated as is discussed in Section 4.2. After the number of conducted traditional retrieval feedbacks reaches 20, Algorithm 1 is executed again.

Figs. 2 and 3 give two examples for retrieval. The blue icons correspond to the deductable retrieval feedback. The green icons correspond to the positive traditional feedbacks. The red icons correspond to the negative traditional feedbacks. The large images are query images; the small images at the left column are deducible retrieval results; the small images at the four right columns are traditional retrieval results. In Fig. 2, the query image is a cervical spine image in sagittal view, and so are the user expected

images. The image is correctly classified. Consequently, the deducible retrieval results contain one image each from the overall labels whose category labels on the “Spine” category do not equal to the “non-” category label. All of the traditional retrieval results are cervical spine images in sagittal view except the last one which is a lumbar spine image in sagittal view.

In the second example shown in Fig. 3, the query image is a cranium image in coronal view while the user expected images are facial cranium images in “others” view, i.e., any view other than coronal, sagittal, and axial. The initial deducible retrieval results contain one image each from the overall labels whose category labels on the “Cranium” category are not the “non-” category label. Since the query image is correctly annotated, the traditional retrieval images are cranium images in coronal view. After the user select the label corresponding to the facial cranium image in “others” view as the feedback to the deducible retrieval, most of the traditional retrieval results become facial cranium image in “others” view. Two of them are not correct because they are incorrectly labelled by SEMI-SECC.

5.3. Evaluations

The first experiment we have conducted is to compare the annotation accuracies between ECOC, which we have implemented based on [5], SECC, and SEMI-SECC. The second column of Table 3 reports the comparison results. The integers and the percentages in “Method” field are the numbers of individual classifiers, i.e., M , and the percentages for the initially ground truthed training samples of all the training samples. Error rate is estimated using the test data only. It is clear from the table that when the M in SECC is comparable to that in ECOC, the error rate of SECC is much less than that of ECOC. We also note that ECOC can finally beat SECC when it uses a substantially larger M . SEMI-SECC methods are also comparable to SECC in performance when the percentage of labelled samples is not less than 5%. We also compare the accuracy of the SECC methods with those of other 12 annotation methods using the same training data and test data (the results of other methods are provided by IMAGECLEF 2005). The highest accuracy is 87.4%; the lowest accuracy rate is 44.3%; the median accuracy is 78.6%. Our method (SECC or SEMI-SECC (not less than 5%)) ranks fourth out of the 13 methods.

Our retrieval method works well when the annotation method to have a high accuracy. It also works well when the annotation fails and the user expected label is among the deducible retrieval results. Thus, the annotation method with the highest accuracy may not be the most suitable one for our retrieval method. Assume that the number of deducible retrieval results is N . Let *related* be the percentage of the queries whose corresponding user expected labels are among the N deducible retrieval results. We use *related* to evaluate how an annotation method is

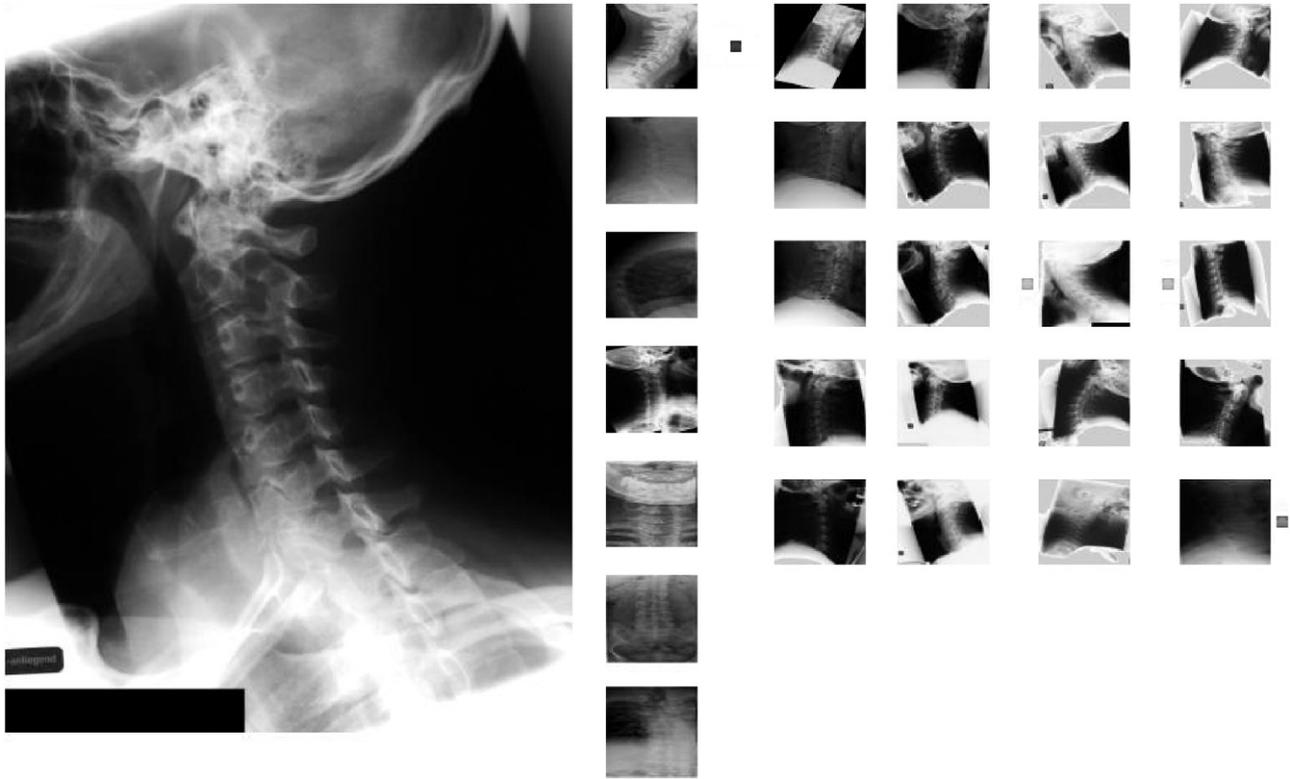


Fig. 2. Example 1: The query image is a cervical spine image in sagittal view; so are the user expected images; the deducible retrieval results contain all the labels in the database whose category labels on the “Spine” category are not the “non-” category label; the traditional retrieval results are correct except the last one.

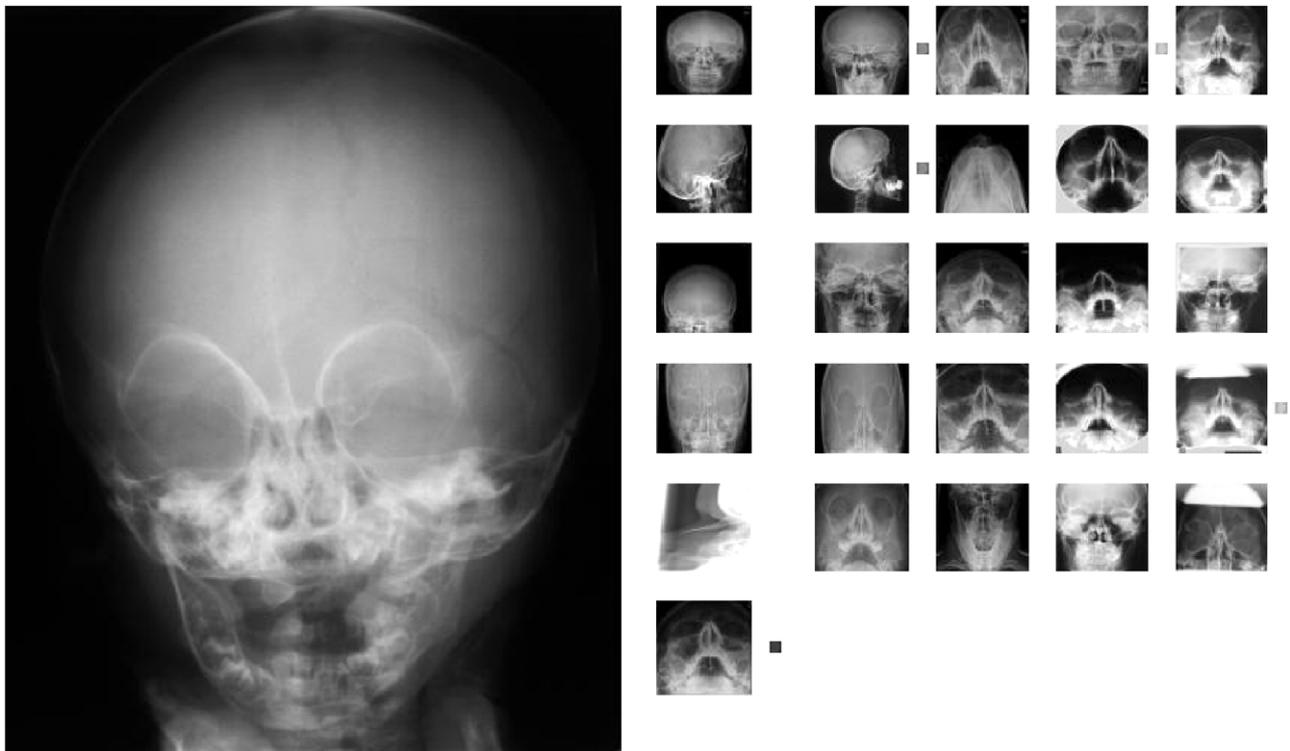


Fig. 3. Example 2: The query image is a cranium image in coronal view; the user expected images are facial cranium images in others view; after one iteration of the deducible retrieval RF, the traditional retrieval results are correct except the first two.

Table 3
Coding method comparisons

Method	Accuracy (%)	Related (%)	Related (%) ^a
SECC (11)	81.3	94.1	93.8
SEMI-SECC (11,2%)	77.1	88.9	88.3
SEMI-SECC (11,5%)	80.7	92.1	91.5
SEMI-SECC (11,10%)	81.1	94.0	93.6
ECOC (10)	67.4	77.3	45.3
ECOC (50)	74.3	83.5	47.1
ECOC (100)	80.5	87.8	49.9
ECOC (200)	84.9	91.6	53.6

The values in parentheses are M and the percentages of initially ground truthed samples. The values in the second and third columns are calculated by considering the ground truth overall label of a query as the correct annotation result of the query. The values in the fourth column are calculated by considering an overall label different from but semantically similar to the ground truth overall label of a query as the correct annotation result of the query.

suitable for the retrieval. As the second experiment, Table 3 reports the *related* values for different annotation methods. Though the accuracy of SECC is less than that of ECOC (200), the *related* of SECC is higher than that of ECOC (200). The reason is that most of the images which are incorrectly annotated still have a correct delegate category. For these images, the user expected label is among the deducible retrieval results when SECC or SEMI-SECC is used.

Since it is possible in our retrieval system that a query image is not exactly but only semantically similar to the user expected images, we also intend to know how the annotation methods perform under this situation. For each test image, we randomly select an overall label different from but semantically similar to the overall label of the query image. This overall label is considered as the correct annotation result of the test image instead of its ground truth overall label. The corresponding *related* values for different annotation methods are reported in the last column of Table 3. It is clear that all the methods except SECC and SEMI-SECC have significant *related* value decreases w.r.t. the corresponding previous results.

The third experiment is to evaluate the percentage for the correctly ground truthed feedback images of all the traditional retrieval feedback images, which is denoted as B , w.r.t. different numbers of conducted traditional retrieval RFs, which is denoted as A . After each 20 traditional retrieval RFs, we record the B values, which are reported in Table 4. In general, B decreases slightly when A increases. The reason is that the larger the A is, the more ground truthed images, the more feedback images which are ground truthed, the smaller the B can be.

Table 4 also reports the retrieval precision w.r.t. different A values. After each 20 RFs of the traditional retrieval, SEMI-SECC is re-trained. Then 100 retrievals are applied under the constraint that the query image has the same

Table 4
Retrieval evaluation results

A	B (%)	C (%)	D
0	63.9	73.9	7.3
20	64.1	75.0	7.1
40	63.5	75.8	7.3
60	63.4	76.6	7.4
80	63.4	77.1	7.6
100	63.2	77.6	7.4

A is the number of conducted traditional retrieval RFs. B is the percentage for the correctly ground truthed feedback images of all the traditional retrieval feedback images. C is the retrieval precision. D is the number (out of 10) of the successful deducible retrievals when the query image has a different label from what the user expects.

label as the user expected labels. It is clear from the table that the precision increases when A increases. The reason is that with the increase of A , the annotation accuracy increases, which leads to the increase of the retrieval precision. In comparison, we also apply the MEDGIFT [2] to the same data set and record a 65.6% precision. This indicates that our retrieval method is promising as there is a significant performance increase.

In order to discover how our retrieval method handles the case when a query image does not have the same label as what a user expects, we have conducted the fourth experiment as follows. First, we randomly select an overall label. Then we query the system using the queries semantically similar to the overall label we have selected. Let D be the number of successful deducible retrievals, whose annotation results for the query images happen to be the corresponding selected ground truth labels, in one section. The number of the queries conducted in each section is 10. We report the average D values for 100 sections w.r.t. different A values in Table 4. It only takes less than two deducible retrieval feedbacks for our system to find the specific label a user expects. Most of the deducible retrievals for the remaining queries in the same section are successful. The experiment also shows that when users modify the labels they expect during one section, D decreases. It is also clear that D is only affected slightly by A , and consequently, is affected slightly by the annotation accuracy.

6. Conclusion

In the paper, we present a novel content based medical image retrieval method which consists of the deducible retrieval and the traditional retrieval. The proposed retrieval method does not require a user to query the system using the exact images he/she expects. The deducible retrieval is to retrieve the label that a user expects while the traditional retrieval is to retrieve the images with the label in the database. The deducible retrieval is achieved using the semi-supervised Semantic Error-Correcting output Codes (SEMI-SECC). RF is used

in both retrieval steps to help identify the user expected label and ground truth the images in the database. We apply the proposed method to IMAGECLEF 2005 and the experimental results clearly show the strength and the promise of the presented methods. Our future work includes how to dynamically better understand the user expect label and how to better ground truth the unlabelled images in the database.

Acknowledgments

This research was funded in part by National Science Foundation (IIS-0535162) and by the intramural research funds of the Lister Hill National Center for Biomedical Communications, the National Library of Medicine, and the National Institutes of Health.

References

- [1] (<http://ir.shef.ac.uk/imageclef2005/>).
- [2] (<http://www.sim.hcuge.ch/medgift/>).
- [3] K. Barnard, P. Duygulu, N.D. Freitas, D. Forsyth, D. Blei, M.I. Jordan, Matching words and pictures, *J. Mach. Learn. Res.* 3 (2003) 1107–1135.
- [4] C.E. Brodley, A.C. Kak, J.G. Dy, C. Shyu, A. Aisen, L. Broderick, Content-based retrieval from medical image databases: a synergy of human interaction, machine learning and computer vision, in: *National Conference on Artificial Intelligence*, 1999, pp. 760–767.
- [5] T. Diettrich, G. Bakiri, Solving multiclass learning problems via error-correcting output codes, *J. Artif. Intell. Res.* 2 (1995) 263–286.
- [6] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
- [7] S.L. Feng, R. Manmatha, V. Lavrenko, Multiple Bernoulli relevance models for image and video annotation, in: *CVPR*, 2004.
- [8] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in: *CVPR*, 2003, pp. 264–271.
- [9] R. Ghani, Using error-correcting codes for text classification, in: *International Conference on Machine Learning*, 2000.
- [10] J. Han, K.N. Ngan, M.J. Li, H.J. Zhang, A memory learning framework for effective image retrieval, *IEEE Trans. Image Process.* 14 (2005) 511–524.
- [11] R. Herbrich, *Learning Kernel Classifiers*, MIT Press, Cambridge, MA, 2002.
- [12] S.C.H. Hoi, M.R. Lyu, A semi-supervised active learning framework for image retrieval, in: *CVPR*, 2005.
- [13] T. Jebara, *Machine Learning Discriminative and Generative*, Kluwer Academic Publishers, Dordrecht, 2004.
- [14] F. Jurie, C. Schmid, Scale-invariant shape features for recognition of object categories, in: *CVPR*, 2004.
- [15] R. Krishnapuram, S. Medasani, S.H. Jung, Y.S. Choi, R. Balasubramaniam, Content-based image retrieval based on a fuzzy approach, *IEEE Trans. Knowl. Data Eng.* 16 (2004) 1185–1199.
- [16] J. Li, N. Allinson, D. Tao, X. Li, Multitraining support vector machine for image retrieval, *IEEE Trans. Image Process.* 15 (2006) 3597–3601.
- [17] B. Liu, W.-S. Lee, P.S. Yu, X.-L. Li, Partially supervised classification of text documents, in: *ICML*, 2002.
- [18] Y. Liu, N. Lazar, W.E. Rothfus, F. Dellaert, A. Moore, J. Schneider, T. Kanade, Semantic based biomedical image indexing and retrieval, in: *Trends and Advances in Content-Based Image and Video Retrieval*, 2004.
- [19] Y. Lu, H. Zhang, W. Liu, C. Hu, Joint semantics and feature based image retrieval using relevance feedback, *IEEE Trans. Multimedia* 3 (2003) 339–347.
- [20] I. Muslea, S. Minton, C. Knoblock, Active + semi-supervised learning = robust multi-view learning, in: *ICML*, 2002.
- [21] S. Rosset, J. Zhu, H. Zou, T. Hastie, A method for inferring label sampling mechanisms in semi-supervised learning, in: *NIPS*, 2004.
- [22] Y. Rui, T.S. Huang, M. Ortega, S. Mehrotra, Relevance feedback: a power tool for interactive content-based image retrieval, *IEEE Trans. Circuits Syst. Video Technol.* 8 (1998) 644–655.
- [23] H.L. Tang, R. Hanka, H.S. Ip, Histological image retrieval based on semantic content analysis, *IEEE Trans. Inf. Technol. Biomed.* 7 (2003) 26–36.
- [24] D. Tao, X. Li, S.J. Maybank, Negative samples analysis in relevance feedback, *IEEE Trans. Know. Data Eng.* 19 (2007) 568–580.
- [25] D. Tao, X. Tang, X. Li, Y. Rui, Kernel direct biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm, *IEEE Trans. Multimedia* 8 (2006) 716–727.
- [26] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval, *IEEE Trans Pattern Anal Mach Intell* 28 (2006) 1088–1099.
- [27] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
- [28] W. Wu, J. Yang, Smartlabel: an object labeling tool using iterated harmonic energy minimization, in: *14th ACM International Conference on Multimedia*, 2006, pp. 891–900.
- [29] J. Yao, Z. Zhang, Object detection in aerial imagery based on enhanced semi-supervised learning, in: *ICCV*, 2005.



Jian Yao received his Ph.D. degree from Computer Science Department of State University of New York at Binghamton. His research areas include computer vision, pattern recognition, machine learning, and data mining. He has published more than 10 papers on top conferences and journals, such as *ICCV*, *CVPR*, *CVIU*, and so on. Currently, he is working for ask.com.



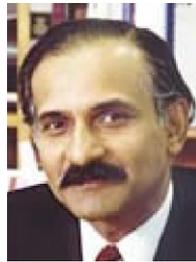
Zhongfei (Mark) Zhang received B.S. (cum laude) in Electronics Engineering, M.S. in Information Science, both from Zhejiang University, Hangzhou, China, and Ph.D. in Computer Science from the University of Massachusetts at Amherst. When he was in the graduate school, he also worked as an Intern student at NEC Research Institute, Inc. at Princeton, NJ, and as a technical consultant at Applied Artificial Intelligence, Inc. (formerly Amerinex Artificial Intelligence, Inc.)

at Amherst, MA. He was a Research Staff Member at the Department of Information Science and Electronics Engineering, Zhejiang University, a Research Scientist at the Center of Excellence for Document Analysis and Recognition (CEDAR), and a Research Assistant Professor at the Department of Computer Science and Engineering, both at SUNY Buffalo. He joined the faculty of Computer Science Department at SUNY Binghamton in the Fall of 1999. He has published over 70 peer-reviewed academic papers in international and national journals and conferences, has served as reviewers or program committee members for many international journals and conferences, and has served as grant review panelists for several governmental and private funding agencies. He is a Senior Member of IEEE, a member of IEEE Computer Society, a member of ACM, and a fellow of the Institute for Student-Centered Learning at Binghamton University. He is an Associate Editor for *Pattern Recognition* published by Elsevier Science.



Sameer Antani is a Staff Scientist with the Lister Hill National Center for Biomedical Communications, an intramural R&D division of the National Library of Medicine, at the National Institutes of Health. He conducts research on various topics in content-based image retrieval, medical multimedia databases, next-generation interactive documents, and advanced multimodal medical document retrieval. He earned his Master of Engineering and Ph.D. in Computer

Science and Engineering from the Pennsylvania State University in 1998 and 2001, respectively. He earned his Bachelors degree in Computer Engineering from the University of Pune, India (Pune Institute of Computer Technology), in 1994. He is a member of the IEEE and the IEEE Computer Society.



George R. Thoma received the B.S. from Swarthmore College, and the M.S. and Ph.D. from the University of Pennsylvania, all in Electrical Engineering. As the Senior Electronics Engineer and Chief of the Communications Engineering Branch of the Lister Hill National Center for Biomedical Communications, a research and development division of the National Library of Medicine, he directs R&D programs in image processing, document image storage on digital

optical disks, automated document image delivery, digital x-ray archiving, and high speed image transmission. He has also conducted research in analog videodiscs, satellite communications and video teleconferencing.



L. Rodney Long is an electronics engineer for the Communications Engineering Branch at the National Library of Medicine, where he has worked since 1990. Prior to his current job, he worked for 14 years in industry as a software developer and as a systems engineer. His research interests are in telecommunications, image processing, and scientific/biomedical databases. He has an M.A. in applied mathematics from the University of Maryland.