Automatic Medical Image Annotation and Retrieval Using SECC

Jian Yao[a,1,*], Sameer Antani[b], Rodney Long[b], George Thoma[b], and Zhongfei Zhang[a]

[a]Department of Computer Science
State University of New York at Binghamton
Binghamton, NY 13902
{jyao,zzhang}@binghamton.edu

[b]National Library of Medicine
National Institutes of Health
Bethesda, MD 20894
{santani,rlong,gthoma}@mail.nih.gov

Abstract

The demand for automatically annotating and retrieving medical images is growing faster than ever. In this paper, we present a novel medical image annotation method based on the proposed Semantic Error-Correcting output Codes (SECC). With this annotation method, we present a new semantic image retrieval method, which exploits the high level semantic similarity. The experimental results on the IMAGECLEF 2005 annotation data set clearly show the strength and the promise of the presented methods.

1 Introduction

Medical images play a central role in patient diagnosis, therapy, surgical planning, medical reference, and medical 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 [5] becomes important.

We consider image annotation as a special classification problem, i.e., classifying a given image into one of the predefined labels. Annotation typically has a large number of possible labels. For example, the number of the different labels for data set from IMAGECLEF 2005 [1] annotation task is 57. Error-Correcting Output Codes (ECOC) [6, 7] is a useful model to solve the classification problems with a large number of possible labels by first solving a set of 2-class classification problems and then combining the classification results from these 2-class classifiers.

Due to the large number of images without text information, content-based medical image retrieval (CBMIR) [4, 9, 10] has received increased attention. Unfortunately, current methods only focus on appearance-based similarity, i.e., the appearance of the retrieved images is similar to that of a query image. There is little semantic information exploited. Among the few efforts which claim to exploit the semantic information, the semantic similarities are defined between different appearances of the same object. We call these kinds of semantic similarities the *low level semantic similarities* and the similarities between different objects the *high level semantic similarities*. The similarity between two images are the similarity between the objects contained by the two images. For example, the similarity between an elbow image in sagittal view and an elbow image in coronal view is the low level semantic similarity while that between an elbow image and a forearm image is the high level semantic similarity.

In this paper, we extend ECOC to a semantic ECOC, which is denoted as SECC. An automatic medical image annotation method is presented based on SECC. With this annotation method, we propose a novel semantic medical image retrieval method, which exploits the high level semantic similarity, in contrast to existing retrieval systems in the literature that are based on the low level semantic similarities. A user may query the database with an image that

*[1]This work was done during participation in the Medical Informatics Training Program of the Lister Hill National Center for Biomedical Communications at the National Library of Medicine, NIH

Appears in Proceedings of 19th International Symposium on Computer-Based Medical Systems (CBMS 2006), June 2006, Salt Lake City, Utah

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)

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

is close to but not exactly what he/she expects. Experimental results not only indicate the promise of the annotation method but also show the strengths of the retrieval methods.

2 Annotation Model

ECOC is used to solve an H-class ($H \gg 2$) classification problem using multiple 2-class classifiers, which are called *individual classifiers*. The procedure to select the individual classifiers is called *coding*. The labels of the original H-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. Each query is first classified by the individual classifiers to form a query code; the overall label whose code is the closest to the query code is selected as the overall label of the query image. Table 1 gives a simple example. There are 4 overall labels: forearm and sagittal, elbow and coronal, foot and axial, and foot and sagittal. 4 individual classifiers are used in an ECOC solution.

The criterion of ECOC coding is that the differences between the codes of different overall labels should be large, which is typically measured using the Hamming distance function. 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. Regarding 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 our method semantically selects the individual classifiers and finds the closest code, i.e., combines the individual classifiers.

2.1 Individual classifiers selection

A typical overall label for IMAGECLEF 2005 annotation data set is "elbow image, sagittal view, plain radiography, and musculoskeletal". We denote each individual part of an overall label as a *category* and the possible values for this category among all the overall labels 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 of a query image and its probability but also the category labels and their probabilities are required to be determined. Since the individual classifiers in the ECOC coding are selected randomly, they seldom contain the latter information. Regarding the example given in Table 1, it is unlikely that an individual classifier solves the classification problem related to one of the three categories exactly. In order to extract such information, we modify ECOC as follows.

First, we define several categories and category labels on 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 category labels corresponding to different aspects of the category, together with a "non-" category label. A sample with a "non-" category label in a category means that the sample does not belong to the category. For 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 labels are non-zero values. For one sample, there is only one correlated category 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. For example, the delegate category of an image from "Label ID 0" in Table 1 is the ARM category.

Then we train one individual classifier for one category. This classifier may be a 2-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.

2.2 Individual classifiers combination

It is clear from above definitions that the SECC coding does not guarantee the differences between the codes of different overall labels to be large. Consequently, the ECOC similarity functions (e.g., the Hamming distance function) are not suitable for SECC. Here we present a probabilistically based similarity function for SECC. Let the number of the individual classifiers be K. Denote the probabilities for a query image to have individual label j of individual

classifier i as q_{ij}. Let $Q = \{q_{ij}\}$. Let a possible code for a query image be $Y = (y^1, y^2, ..., y^K)$ and the code of the overall label o be $G_o = (g_o^1, g_o^2, ..., g_o^K)$. We maximize the joint probability of G_o and Y given Q to find the overall label of the query image:

$$Max_{o,Y}\ P(G_o, Y|Q) = P(G_o|Y, Q) \times P(Y|Q) \quad (1)$$

where $P(Y|Q)$ is the probability that the individual classification results are y^i given q_{ij}. Different individual classifiers are trained independently. Consequently, it is possible that for some Y, there are more than one y^j for correlated categories which are not 0. Recall that this is in conflict with the fact that there is only one delegate category. Hence, the corresponding $P(Y|Q)$ is set to 0. For the other cases, $P(Y|Q)$ is set to the multiplication of the probabilities that individual classification labels are correct, i.e., q_{iy_i}. Let y^{C_i} be the y^i of the correlated categories. $P(Y|Q)$ is defined as follows:

$$P(Y|Q) = \begin{cases} 0, |\{y^{C_i}, y^{C_i} \neq 0\}| \neq 1 \\ \prod_{i=1}^{K} q_{iy_i}, |\{y^{C_i}, y^{C_i} \neq 0\}| = 1 \end{cases} \quad (2)$$

$P(G_o|Y, Q)$ in Equation 1 is the probability of the event that a query code Y with the probability set Q happens to be the ground truth code G_o. To simplify the computation, we let $P(G_o|Y, Q) = P(G_o|Y)$. Let $D_o = |\{j, g_o^j \neq y^j\}|$. We then define $P(G_o|Y)$ as follows:

$$P(G_o|Y) = \begin{cases} 0,\ D_o \geq T \\ P(\{(j, g_o^j), g_o^j \neq y^j\}|\{(j, g_o^j), g_o^j = y^j\}),\ D_o < T \end{cases} \quad (3)$$

The conditional probability in the right hand side of Equation 3 is the probability of the event that when a query code contains part of a ground truth code, the remaining part of the query code happens to be the remaining part of the ground truth code. In order to focus the attention on the query codes that do not differ substantially from the codes of possible overall labels, we introduce a threshold T. If the code of an overall label differs from the query code by at least T 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.), $P(\{(j, g_o^j), g_o^j \neq y^j\}|\{(j, g_o^j), g_o^j = y^j\})$ can be estimated using the training samples. For example, referring to the example in Table 1, assume that Label 0 has 20 training samples and Label 1 has 30 training samples. Since only Label 0 and Label 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 Retrieval Model

CBMIR is concerned with retrieving images in a database that are similar to a query image in content. A key difference between the existing retrieval systems, which are denoted as the *traditional retrieval* in the following text, and the semantic retrieval method presented in this paper is that in the former, query images are visually similar to the images an user interests while in the latter, query images only need to be similar at a high semantic level to the images an user interests. For example, in our retrieval system, an upper arm image can be used to retrieve hand images. This is unlikely to happen in the existing retrieval systems.

Since our semantic retrieval focuses on the similarities among different objects at the high semantic level, we must define these similarities in advance. The challenge is in addressing the subjective nature of human semantic interpretation of images. For example, the same similarity may be defined between different views of the same object, or between different parts of the same object, or between different objects. In our semantic retrieval method developed for the IMAGECLEF 2005 annotation data set, the similarity between different objects is defined through the similarity between their overall labels. In the current version of our semantic retrieval prototype method, the similarity between two objects is either 0 (not similar) or 1 (similar). Two overall labels are similar if their delegate categories are the same.

For a query image, we first apply the SECC annotation method to determine the individual labels and their probabilities. The overall label is then determined using the method presented in Section 2.2. Based on the similarities defined above, all the overall labels which are semantically similar to the classified overall label of the query image are extracted. The retrieved images of our semantic retrieval consist of one randomly selected image from each of these overall labels. Consequently, our semantic retrieval retrieves labels instead of images.

Since the ultimate goal of the image retrieval is to retrieve images instead of labels, our semantic retrieval must be combined with traditional retrieval as follows. First, our semantic retrieval is applied to determine the overall labels with which a user may expect to retrieve images. It is either the annotation result or a user selected overall label from the results of our semantic retrieval. A traditional retrieval method may then be applied to retrieve the images in a database with this overall label. Since our semantic retrieval is subjective, relevance feedback (RF) [8] may be used to determine the specific overall label which the user expects. Our future work will focus on how to combine the RF to effectively retrieve images.

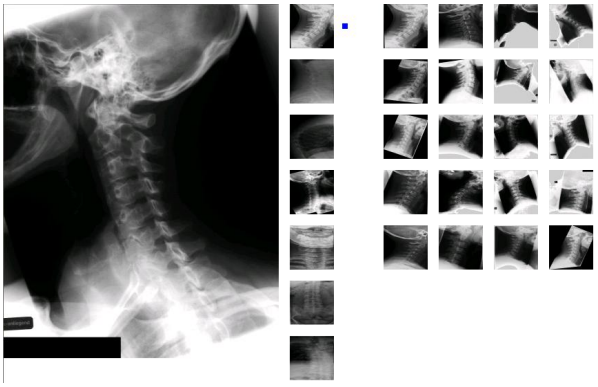

Figure 1. Retrieval example #1: The query image is a cervical spine image in sagittal view; the user expected images is also cervical spine image in sagittal view; the semantic retrieval results contain all the labels related to spine images; the traditional retrieval results are correct.

4 Evaluations

We use the image set from IMAGECLEF 2005 [1] annotation task to evaluate our methods. 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 to 2563 training images.

We define 11 categories for the data set: CRANIUM(C), SPINE(C), ARM(C), LEG(C), VIEW(I), RADIOGRAPHY(I), FUNCTION(I), CHEST(C), ABDOMEN(C), PELVIS(C), and BREAST(C), where C or I represents a correlated category or an independent category. Each category has 2 to 6 labels.

Each image is first normalized into size 16×16. We have compared three kinds of features: intensity feature, Harr wavelet feature, and Garbor wavelet feature. Combining the computation efficiency and retrieval effectiveness, we select the intensity as the feature for the retrieval.

We test our annotation and retrieval models under the platform of Pentium IV 2GHz CPU with 512M memory. The training procedure takes about 4 hours using the 9000 training images. The annotation for the 1000 test images takes about 4.5 minutes. A test image is successfully annotated if the annotated overall label equals to the ground truth overall label. The retrieval (without feedback) for a test image takes about 0.3 second.

Figure 1 and Figure 2 give two examples for retrieval. The large images are the query images; the small images at the left column are imaginary retrieval results; the small images at the 4 right columns are our semantic retrieval results. In Figure 1, the query image is a cervical spine image in sagittal view, and so are the user expected images. The image is correctly classified. Consequently, our semantic retrieval results contain one image each from the overall labels whose category labels on "SPINE" category do not equal to the "non-" category label.

In the second example shown in Figure 2, 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 semantic retrieval results contain one image each from the overall labels whose category labels on "CRANIUM" category do not equal to the "non-" category label. Since the query image is correctly annotated, our semantic 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 semantic retrieval, most of the traditional retrieval images become facial cranium image in others view.

Coding Method	Error Rate
SECC (11)	18.7%
ECOC (10)	32.6%
ECOC (50)	25.7%
ECOC (100)	19.5%
ECOC (200)	15.1%

Table 2. Comparisons between ECOC and SECC. The values in the parentheses are the numbers of the used individual classifiers.

4.1 Annotation evaluations

Table 2 documents the comparisons between SECC and ECOC which we have implemented based on [6]. The numbers in "Coding Method" field are the numbers of the individual classifiers, i.e., K. It is clear from the Table that when the number of the individual classifiers in SECC is comparable with 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 large K (e.g., 200).

We also compare the accuracy of the SECC method with 12 other annotation methods using the same training data and test data (the results of other methods are provided by IMAGECLEF 2005 [2]). The lowest error rate is 12.6%; the highest error rate is 55.7%; the median error rate is 21.4%. Our method (18.7%) ranks fourth out of the 13 methods.

4.2 Retrieval evaluations

In order to evaluate our retrieval method, several retrieval methods are built for comparisons. They all follow the two

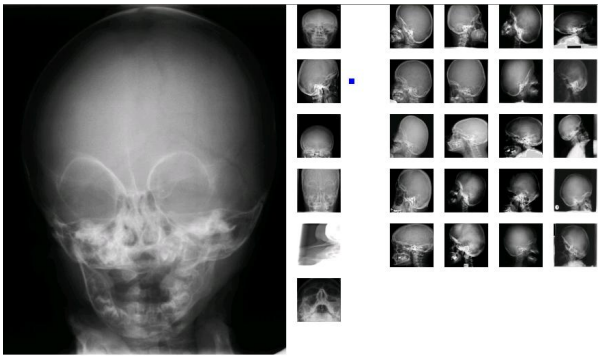

Figure 2. Retrieval example 2: The query image is a cranium image in coronal view; the user expected images are facial cranium images in others view; after user select its expected type among semantic retrieval results, the traditional final retrieval results are correct.

level retrieval structure we have mentioned above, i.e., a semantic retrieval followed by a traditional retrieval. The only difference is in the semantic retrieval level, where different retrieval methods use different annotation methods. The semantic retrieval methods which use the ECOC annotation methods return the 5 overall labels which have the maximum similarities to the query image. The semantic retrieval method which uses the SECC annotation method returns the 5 overall labels which are high level similar to the query image and have the maximum similarities to the query image. The images in the test database are query images. An query is successful if the user expected label is among the semantic retrieval results. In the first example, the query image has the same overall label as the user expected label. The precisions of different retrieval methods are documented in Table 3. It is clear that the precision of the SECC retrieval is higher than that of the ECOC (200) retrieval, though the accuracy of the SECC annotation is less than that of the ECOC (200) annotation as we have shown in Table 2. The reason is that most of the images which are incorrectly annotated still have a correct delegate category. Since the results of our semantic retrieval, i.e., the SECC retrieval, only depend on the classified delegate category, the SECC retrieval results are still correct for those images. We have also applied MEDGIFT [3], which is a traditional retrieval method, to the same data set. The precision is 65.6%. This indicates that our retrieval method is promising as there is a significant performance increase.

In Experiment 2, the user expected images are high level similar but not equal to the ground truth overall label of the query image. Table 3 reports the corresponding precisions. It is clear that all the methods except the SECC method have

Method	Exp. 1	Exp. 2	Exp. 3
SECC (11)	94.1%	93.8%	9.2%
ECOC (10)	77.3%	45.3%	10.3%
ECOC (50)	83.5%	47.1%	8.8%
ECOC (100)	87.8%	49.9%	9.6%
ECOC (200)	91.6%	53.6%	11.2%
MEDGIFT	65.6%	27.3%	2.1%

Table 3. Comparisons between different retrieval methods. The percentages are the precisions. In Experiment 1, the query image has the same overall label as the user expected label; in Experiment 2, the query image is high level similar to the user expected images; in Experiment 3, the query image and the user expected images are independent.

a significant precision decrease w.r.t. Experiment 1. In Experiment 3, the user expected label is independent to the query image. Table 3 reports the corresponding precisions, where all the methods have similar low precisions. The reason is that when the query image is independent to the user expected images, any retrieval is equal to a random retrieval.

5 Conclusions

In this paper, we present a novel medical image annotation method based on SECC. With this annotation method, we present a novel image retrieval method. The experimental results on IMAGECLEF [1] annotation data set clearly show the strength of these methods.

6 Acknowledgement

This research was supported [in part] by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

References

[1] http://ir.shef.ac.uk/imageclef2005/.

[2] http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef05_aat_results.html.

[3] http://www.sim.hcuge.ch/medgift/.

[4] C. E. Brodley, A. C. Kak, J. G. Dy, C. Shyu, A. Aisen, and 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.

[5] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *Computer Vision and Pattern Recognition*, 2005.

[6] T. Diettrich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[7] R. Ghani. Using error-correcting codes for text classification. In *International Conference on Machine Learning*, 2000.

[8] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8:644–655, 1998.

[9] H. Shao, W.-c. Cui, and H. Zhao. Medical image retrieval based on visual contents and text information. In *International Conference on Systems, Man and Cybernetics*, 2004.

[10] H. L. Tang, R. Hanka, and H. S. Ip. Histological image retrieval based on semantic content analysis. *IEEE Trans. On Information Technology in Biomedicine*, 7:26–36, 2003.