AUTOMATIC MEDICAL IMAGE ANNOTATION AND RETRIEVAL USING SEMI-SECC

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

[a]Department of Computer Science
State University of New York at Binghamton
Binghamton, NY 13902

[b]National Library of Medicine
National Institutes of Health
Bethesda, MD 20894

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 based on SEMI-supervised Semantic Error-Correcting output Codes (SEMI-SECC). The experimental results on IMAGECLEF 2005 [1] 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 [4] becomes important.

Due to the large number of the images without text information, content-based medical image retrieval (CBMIR) [2, 3] 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 intra-object similarity while the semantic similarity between a hand image and an upper-arm image is inter-object similarity.

The problem addressed in this paper is a special medical image retrieval problem. Compared with the general medical image retrieval 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.

The rest of the paper is organized as follows: the annotation method is presented in Section 2; the retrieval method is presented in Section 3; the evaluations of the annotation method and the retrieval method using the data set from IMAGECLEF 2005 [1] annotation task are given in Section 4; finally, the conclusion is made in Section 5.

2. ANNOTATION MODEL

2.1. Error-Correcting Output Codes (ECOC)

ECOC [5, 6] 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. Table 1 gives a simple example, where 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 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

[1]Part of 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

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

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 our method selects the individual classifiers and finds the closest code, i.e., combines the individual classifiers.

2.2. 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 revise ECOC to SECC.

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 other category labels are non-zero. 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 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. 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.

2.3. 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 $Y = (y^1, y^2, ..., y^M)$ and the code of overall label o as $G_o = (g_o^1, g_o^2, ..., g_o^M)$. We maximize the joint probability of G_o and Y given Q_i to find the overall label of the query image:

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

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., $q_i^{jy_j{}'}s$. Let $y^{C_j{}'}s$ be the $y^{j'}s$ for the correlated categories. We then define $P(Y|Q_i)$ as follows:

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

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

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

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 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.), $P(\{(j, g_o^j), g_o^j \neq y^j\}|\{(j, g_o^j), g_o^j = y^j\})$ can be estimated using the training samples. For example, referring to the example in Table 1, assume that Label 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:

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

2.4. 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 ESL framework presented in [7]. The ESL framework is probabilistically guaranteed to have the accuracy increased when the number of iterations increases. The SEMI-SECC learning procedure is summarized in Algorithm 1.

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. Re-train the individual classifiers.
5. If certain stop criteria meet, stop. Otherwise, goto step 3.

3. RETRIEVAL MODEL

Image retrieval concerns with retrieving images in a database that are similar to a query image in content. We call the retrieval systems that use appearance-based or low level semantic similarities as the *traditional retrieval* systems and those that use high level semantic similarities as the *imaginary retrieval* systems. Existing retrieval methods in the literature are all traditional retrieval. A big difference between the traditional retrieval and the imaginary retrieval is that in the traditional retrieval system, query images are the same as the user interested images while in the imaginary retrieval system, query images are high level semantical similar to the user interested images.

Since the imaginary retrieval focuses on the similarities among different objects, we must define such similarities in advance. Unfortunately, such similarities are subjective. For example, for the same similarity, it may be defined for between different views of the same object, or for different views of different objects that look similar, or different parts of the same object. In the imaginary retrieval system developed for the IMAGECLEF 2005 annotation data set, the similarity between different objects is defined through the similarity between different overall labels. In the current version of the imaginary retrieval prototype system, the similarity between any two objects is either 0 (not similar) or 1 (similar). Two overall labels are similar if there exist two corresponding individual labels between them: (1) they are labels of correlated categories; and (2) they are valuable labels.

For a query image, we first apply the SEMI-SECC annotation method to determine the individual labels and their probabilities. The overall label is then determined using the method presented in Section 2.3. Based on the similarities defined above, all the overall labels which are similar to the overall label of the query image are extracted. The imaginary retrieval images are randomly selected images with each of these overall labels. Using the four overall labels in the example discussed in Section 2.2, if a query image is annotated to have a "forearm" label, the retrieved images are those either with a forearm label or with an elbow label.

The imaginary retrieval may be combined with the traditional retrieval to develop a more sophisticated, hierarchical retrieval system. For example, the imaginary retrieval may first be applied to determine the overall label with which a user expects to retrieve images. A traditional retrieval system may then be applied to actually retrieve the images in a database with this overall label.

4. EVALUATIONS

The data set we use to evaluate our methods is IMAGECLEF [1] 2005 annotation data set. All the images are X-Ray images, which include 9000 training images and 1000 test images. The images can be categorized into 57 classes. We define 11 categories for the data set.

The first experiment we have conducted is to compare the annotation accuracy between the ECOC methods, which we have implemented based on [5], the SECC method, and the SEMI-SECC methods. The second column of Table 2 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 er-

Method	Accuracy	*Related*	*Related**
SECC (11)	81.3%	**94.1%**	**93.8%**
SEMI-SECC (11,2%)	75.6%	87.9%	88.0%
SEMI-SECC (11,5%)	78.1%	91.5%	91.5%
SEMI-SECC (11,10%)	79.4%	93.3%	93.4%
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%

Table 2. Coding methods comparisons. 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.

ror 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 lowest error rate is 12.6%; the highest error rate is 55.7%; the median error rate is 21.4%. Our method (SECC or SEMI-SECC (not less than 5%)) ranks fourth out of the 13 methods.

Our retrieval method not only requires the annotation method to have a high accuracy, but also requires that when the annotation fails, 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 suitable for the retrieval. As the second experiment, Table 2 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 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 2. It is clear that all the methods except SEMI-SECC have significant *related* value decreases w.r.t. the corresponding previous results.

5. CONCLUSIONS

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 based on SEMI-supervised Semantic Error-Correcting output Codes (SEMI-SECC). The experimental results on IMAGECLEF 2005 annotation data set clearly show the strength and the promise of the presented 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).

7. REFERENCES

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

[2] S. Antani, R. Long, and G. Thoma. Content-based image retrieval for large biomedical image archives. In *Proceedings of 11th World Congress on Medical Informatics (MEDINFO)*, 2004.

[3] 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*, pages 760–767, 1999.

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

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

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

[7] J. Yao and Z. Zhang. Object detection in aerial imagery based on enhanced semi-supervised learning. In *ICCV*, 2005.