Automatic multi-label annotation of abdominal CT images using CBIR

Zhiyun Xue, Sameer Antani, L. Rodney Long, George R. Thoma
National Library of Medicine, National Institutes of Health, Bethesda, MD

ABSTRACT

We present a technique to annotate multiple organs shown in 2-D abdominal/pelvic CT images using CBIR. This annotation task is motivated by our research interests in visual question-answering (VQA). We aim to apply results from this effort in Open-i^{SM}, a multimodal biomedical search engine developed by the National Library of Medicine (NLM). Understanding visual content of biomedical images is a necessary step for VQA. Though sufficient annotational information about an image may be available in related textual metadata, not all may be useful as descriptive tags, particularly for anatomy on the image. In this paper, we develop and evaluate a multi-label image annotation method using CBIR. We evaluate our method on two 2-D CT image datasets we generated from 3-D volumetric data obtained from a multi-organ segmentation challenge hosted in MICCAI 2015. Shape and spatial layout information is used to encode visual characteristics of the anatomy. We adapt a weighted voting scheme to assign multiple labels to the query image by combining the labels of the images identified as similar by the method. Key parameters that may affect the annotation performance, such as the number of images used in the label voting and the threshold for excluding labels that have low weights, are studied. The method proposes a coarse-to-fine retrieval strategy which integrates the classification with the nearest-neighbor search. Results from our evaluation (using the MICCAI CT image datasets as well as figures from Open-i) are presented.

Keywords: multi-label image annotation, content-based image retrieval, abdominal CT

1. INTRODUCTION

The National Library of Medicine's (NLM) Open-i^{SM} multimodal biomedical information search engine indexes Open-Access biomedical literature, medical cases, and biomedical images. There are figures and images in the collection that are abdominal CT images. These images are often individual 2-D axial slice images which are selected from the original DICOM formatted volumetric data and converted to a regular JPEG or PNG formatted image with intensity range [0,255]. The appearance of abdominal CT slices is quite diverse with respect to the anatomical structure. To automatically annotate the anatomical structures contained in an abdominal CT figure, in addition to the visual content of the image, we could also use the relevant textual metadata, e.g., the figure caption and mention in the article, or the description in a medical case. In this paper, we present a method to assign organ labels to an image based on using the visual information alone. The question to which we hope to be able to answer by the computer is: *given an axial abdominal CT image only, which organs are shown in this image?* In prior work, we have reported our research results in automatically annotating/classifying images based on image type, body segment, and view. Organ annotation, the task in this paper, is a further/finer level of annotation. Unlike the annotation with respect to image modality, body segment, or image view which is single-label annotation and is usually treated as a classification problem, our organ annotation task is a multi-label annotation task which is more challenging. In addition to this, we also face challenges such as, 1) there is a limited amount of ground truth data (annotated images); 2) there are relatively large numbers of organ types which need to be identified; 3) some images only have subtle differences which are confined to small areas; 4) the contents in some images may vary significantly across the images (for example, the difference between the slices that are close to the thoracic cavity and the slices that are close to the pelvic cavity). Because of these challenges, instead of considering multi-label annotation as a classification problem, though some approaches have used deep learning for this [1] and other techniques use visual words to annotate tiled image subregions [2], we choose to tackle this task using content-based image retrieval (CBIR) techniques. CBIR technology is used to find images similar in their visual characteristics. The CBIR technique has been previously used for automatic image annotation [3-5]. The general idea is to use CBIR to identify/retrieve a set of labeled images that are visually similar to a given unlabeled query image, and to annotate this unlabeled image using labels/annotations from visually similar labeled images. To the best of our knowledge, there is no work reported in the literature which addresses automatic annotation of multiple organs in abdominal CT images. The datasets we used to evaluate our method are two 2-D CT image datasets we generated from 3-D volumetric data obtained from a multi-organ segmentation challenge hosted in MICCAI 2015. The insights gained by this work will help us in developing algorithms for visual question-answering (VQA). The goal of VQA is to respond to textual and/or visual queries with relevant images. It is therefore critical that the images are fully annotated to support understanding of the visual content to enable meaningful responses to multimodal queries posed to the search engine.

2. METHOD

2.1 MICCAI Volumetric CT Datasets

The 2015 MICCAI conference organized a “Multi-Atlas Labeling beyond the Cranial Vault” challenge [6]. The main goal of the challenge was to evaluate the performance of automated segmentation methods on two clinically acquired 3-D CT datasets. One set contained 50 abdomen CTs in which 13 abdominal organs (right/left adrenal glands, aorta, esophagus, gall bladder, right/left kidney, liver, pancreas, splenic/portal veins, spleen, stomach, and vena cava) are manually delineated. The other set contained 50 pelvic CTs of cervical cancer patients in which 4 pelvic organs (the uterus, bowel, bladder and rectum) are manually delineated. In both sets, the manual markings of organs in 30 scans are made publicly available (provided as the training data for the segmentation methods). We use these labeled 30 CT scans in each set to develop and quantitatively evaluate our automatic image annotation (specifically, organ annotation) approach.

2.2 Extraction and processing of the 2-D CT slice images

We first extract all the 2-D slices in the axial view from the 3-D CT scans. The images needed by our task should be 2-D axial view CT images in the regular image format (like those figure images in Open-i), which has pixel intensity in the range of (0 – 255). However, the pixel intensity of the CT scans is in Hounsfield Units (HU) and has a much larger range, for example, between -3024 to 3095 for some scans. To convert the intensity from HU to the regular range [0, 255], we apply the windowing technique which takes two parameters: window width and window level. We set the window width to be 400 and the window level to 60 (based on the window presets given in [7]). We also resample each slice to the isotropic grid to get the right image aspect ratio using the slice spacing information. For some CT scans, the slices also need to be rotated/flipped in order to obtain a standard orientation where the back of the body is shown on the lower part of the image. Sometimes, there are regions other than the body appearing in the image. These non-body regions are removed automatically by assuming the largest region in the slice is the body region. The images are then cropped to the size of the bounding box of the body region. Note that not all the slices have labels. In order to quantitatively evaluate the annotation performance, we use a subset of this all-slice dataset which excludes the slices that do not have labels. The resulting labeled abdomen slice dataset consists of 2150 images and the labeled pelvic slice dataset consists of 2518 images. Figure 1 and 2 provide several example images in which organs are marked in the abdominal dataset and the pelvic dataset, respectively.

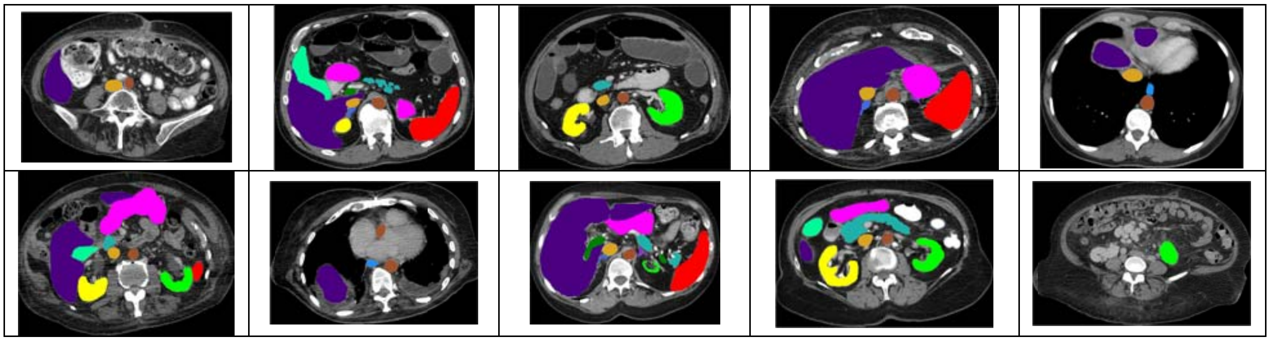


Figure 1. Examples of marked slice images from abdominal CT scans

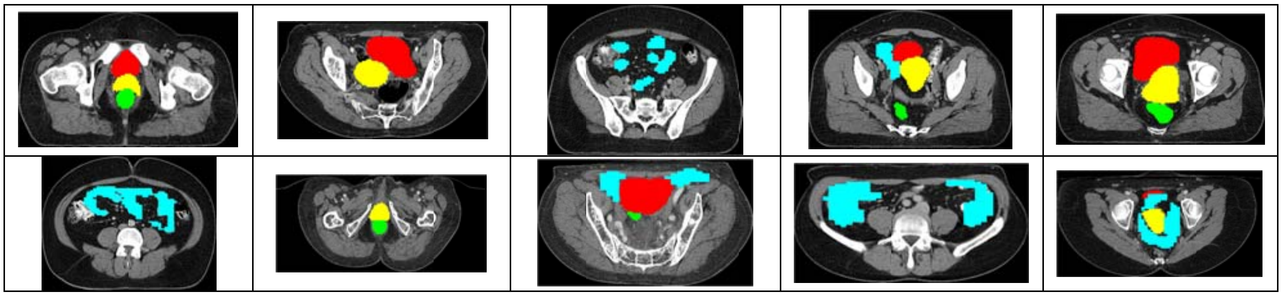


Figure 2. Examples of marked slice images from cervix CT scans

2.3 Image retrieval method

For CBIR, images need to be represented with visually descriptive features. Various image features reported in the literature can be broadly categorized into four groups: color features, texture features, shape features, and spatial relation features. A good survey on features used in CBIR can be found in [8]. We use and evaluate three features in our application: PHOG, SPCEDD, and GIST. PHOG (Pyramid of Histograms of Orientated Gradients) [9] is mainly inspired by two sources: the histogram of oriented gradients (HOG) descriptor and the image pyramid representation. We use the PHOG descriptor because it represents an image by two aspects: its local shape and the spatial layout of the shape. We expect it to be useful in capturing the characteristics of the CT slice images which include different organs with respect to their shapes and locations. For PHOG, the local shape is represented by a distribution/histogram of edge orientations within an image sub-region and the spatial layout of shapes is represented by dividing the image into a sequence of increasingly finer spatial grids (i.e., a spatial pyramid). We use SPCEDD (spatial pyramid of color and edge directivity descriptor) [10] in our experiments because CEDD [11] is one of the most effective features in our work for image modality classification, and it takes edge distribution into consideration. Similar to PHOG, SPCEDD is a concatenation of all the CEDD vectors computed for each grid cell at each pyramid resolution level. The length of the SPCEDD feature is defined as $144\sum_{\mathrm{i}=0}^{\mathrm{L}} 4^{\mathrm{i}}$ (where L denotes the pyramid level), as the length of CEDD vector is 144. GIST was initially proposed by Oliva and Torralba [12] for scene recognition in which the region segmentation step is circumvented. The GIST descriptor aims to extract the spectral and coarsely localized information in the image which can be used to represent/estimate the dominant spatial properties/structure of the scene. The length of GIST vector is equal to the product of the number of sub images, the number of scales and the number of orientations for Gabor filters. We use histogram intersection to measure the similarity between two feature vectors.

2.4 Image label combination

Given an image to be annotated, i.e., the query image, we first obtain n most similar images using CBIR. Next, we need to combine and propagate labels of the top returned images to get the labels for the query image. The method we use to combine the labels is adapted from the method proposed and used for single-label annotation in [3]. A weighted voting scheme is used that gives more weight to the labels of the images that are more similar to the query. Assuming the similarity values of the top K returned images $I_1, I_2, \cdots, I_K$ are $s_1, s_2, \cdots, s_K$, respectively, and $s_1 > s_2 > \cdots > s_K$, the weighted vote v_i for the i^{th} most similar image is defined as:

$$v_i = (s_i + \epsilon)/(s_1 + \epsilon) \tag{1}$$

where s_i is the similarity value of the ith most similar image and ϵ is a small value which is used to avoid divisions by zero. Let denote M as the number of overall organ types/labels in the dataset (for example, 13 for the abdominal dataset) and $(L_1, L_2, \cdots, L_M)$ as the collection of the labels. The weighted vote V_j for label L_j is defined as:

$$V_j = \frac{1}{K}\sum_{i=1}^{K} \lambda_{ji} v_i \text{ , and } \lambda_{ji} = \begin{cases} 1 & \text{if } L_j \text{ is one of the labels of } I_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The final set of labels assigned to the query image consists of those labels whose corresponding weight vote value of each label is larger than a threshold (for example 0.5). That is, annotating the query image with label L_j $(j = 1,2, \ldots, M)$ if $V_j \geq T$ where T is a threshold value smaller than 1.

2.5 Evaluation measures

To assess the performance of annotation, we compare the extracted labels with the ground truth labels for each query image in a test set. Specifically, we calculate the following values:

- N_{gt_labels}: the overall number of ground truth labels for all the query images;
- N_{system_labels}: the overall number of extracted labels for all the query images;
- $N_{missing\_labels}$: the overall number of missing labels (a label that is one of the ground truth labels but not one of the extracted labels for a query image);
- N_{extra_labels}: the overall number of extra labels (a label that is one of the extracted labels but not one of the ground truth labels for a query image);
- $N_{query\_with\_missing\_labels}$: the number of query images that have at least one missing label;
- $N_{query\_with\_extra\_labels}$: the number of query images that have at least one extra label;
- $N_{query\_with\_missing\_and\_extra\_labels}$: the number of query images that have at least one missing label and one extra label.

All above values are then divided by the total number of query images (N_{query}). Smaller values of $N_{missing\_labels}$, N_{extra_labels}, $N_{query\_with\_missing\_labels}$, $N_{query\_with\_extra\_labels}$, $N_{query\_with\_missing\_and\_extra\_labels}$ with respect to the value of N_{query} indicate better performance. Two important parameters that may affect the annotation performance are K, the number of retrieved images, and T, the threshold value for excluding labels that have low weights.

2.6 Coarse-to-fine image retrieval method

Generally speaking, the type of organs and the number of organs in an image slice are related to the location of the slice in the body (along the z-direction). We propose a coarse-to-fine retrieval strategy aiming to approximate the slice location and use it in retrieval. This strategy consists of two steps. It first coarsely classifies the query image into one body segment and then finds the most similar images among those images which are in the same body segment. This is the approach we use in the experiments for improving system performance. For example, in Experiment 2 (Section 3.2), for the abdominal dataset, the slices ordered from the bottom to top in the z-direction are broadly divided into 4 consecutive body groups based on whether the images contain certain organs. The features of the slices in each group are extracted and used to train a supervised classifier. The data used to train the classifier consists of all the slices in the retrieval database (none from the query database). Additional details on this method are provided in Section 3.2.

3. EXPERIMENAL RESULTS AND DISCUSSION

To analyze and evaluate the performance of the proposed annotation method, we carry out several experiments using different query image databases and retrieval image databases. In Experiment 1, the query image database and the retrieval image database are identical and contain all of the labeled slices from the abdomen CT or the pelvic CT datasets; for each query, the query image is excluded from the retrieval database. The PHOG feature and nearest-neighbor (NN) search strategy on the entire dataset are used in Experiment 1. The results demonstrate the high performance of this approach (PHOG + NN) for this experiment. In Experiment 2, the images in the query set and the retrieval set are from different patients. As expected, the performance of the approach (PHOG + NN) is worse than that in Experiment 1. To improve the performance, we examine the effect of using different parameters for PHOG, and we test additional features such as SPCEDD and GIST. We also compare this method to the proposed coarse-to-fine retrieval method. Experiment 3 provides preliminary results on the performance of this technique on figures in Open-i which are more diverse and more dissimilar to the labeled dataset.

3.1 Experiment 1

Figure 3 shows the retrieval and annotation results for one example abdominal slice image. For this case, the number of retrieved images (K) is set to 10, and the threshold value (T) for excluding labels that have low weights is set to 50%. The ground truth labels for this query image are right kidney, left kidney, gallbladder, liver, aorta, inferior vena cava, and pancreas. The annotation labels output by the system have one missing label (gallbladder). Although 5 images out of 10 have gallbladders, the overall voting weight for label gallbladders is less than the threshold T (50%) because the voting also considers the similarity ordering of the images, and those 5 images are not among the top 5 returned images. As shown in Figure 3, this case demonstrates the challenges and complexity of our task, such as: there are subtle differences between images, the regions that need to be labeled may be quite small and hard to identify, the number of regions to be annotated is relatively high, and the shape of regions changes across the images. For this parameter setting (K = 10 and T = 0.5), on average, each query image has 5 or 6 labels, the system's annotation for each image contains 5 or 6 labels, and the number of missing labels or extra labels for each image is either 0 or 1. We also examine the annotation performance with respect to the effects of parameters K and T. Figure 4(a) shows the values of annotation performance measures plotted against the increasing value of the number of retrieved images K (from 1 to 25) for a given value of T (= 50%) for abdomen data. Generally speaking, the performance gets worse with increasing K, which indicates more dissimilar images are taken into consideration for annotation. It also implies that the retrieval performance is very good, as the higher the rank of the returned image, the more similar the image is. The figure also demonstrates the tradeoff between the number of missing labels and the number of extra labels. Figure 4(b) plots the values of annotation performance measures with the increasing value of the threshold T for the label voting (from 0.1 to 0.8) for a given value of K (= 10) for abdomen data. As shown by Figure 4(b), selecting the value about 0.5 for T is a good choice, which may be consistent with our intuition. We carry out the same experiment for the cervix dataset also. For most of the cases, the annotations are accurate. The above analysis on parameters K and T based on the abdomen data is also applicable to the cervix data. Compared to the results of abdominal data, the performance on the cervical data is better. It is probably because of two factors: 1) the number of labels to annotate in cervix data is fewer than that of abdomen data (4 vs. 13); 2) there are more variance and complexity shown in the images of abdomen data than in the images of cervix data.

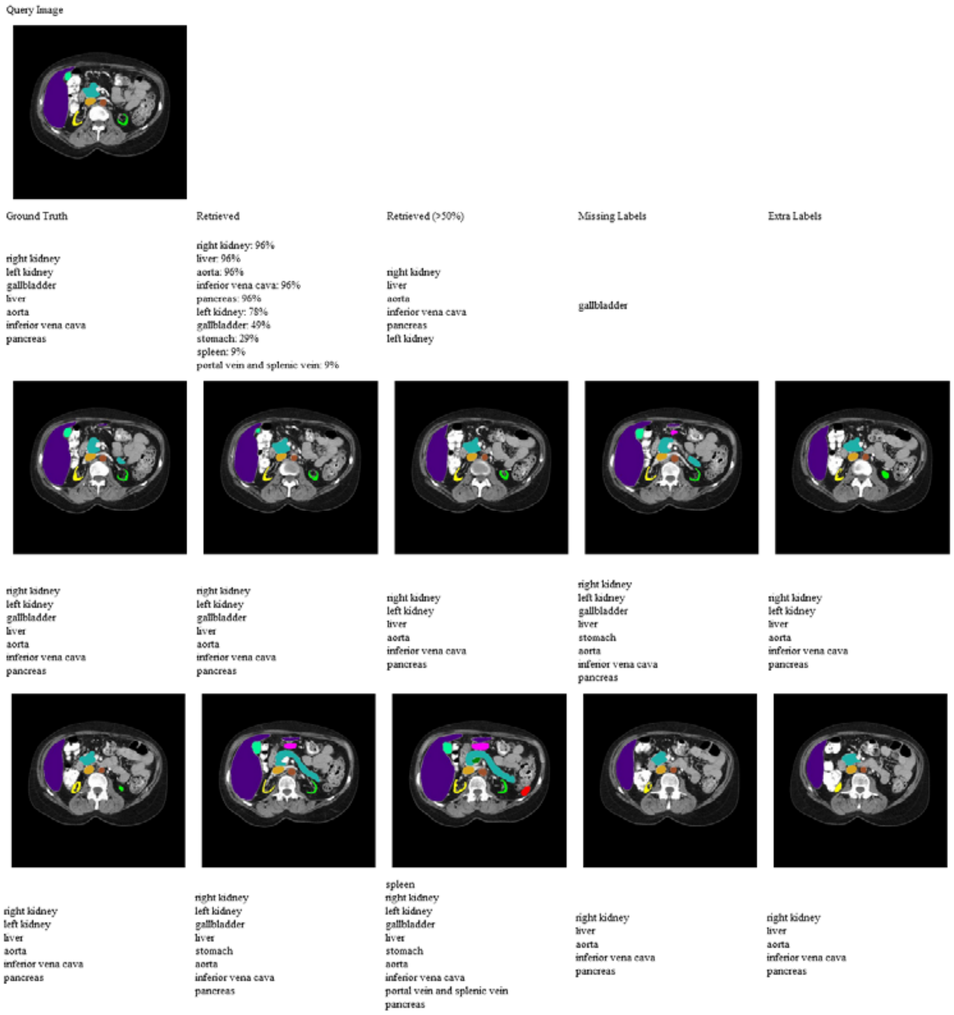
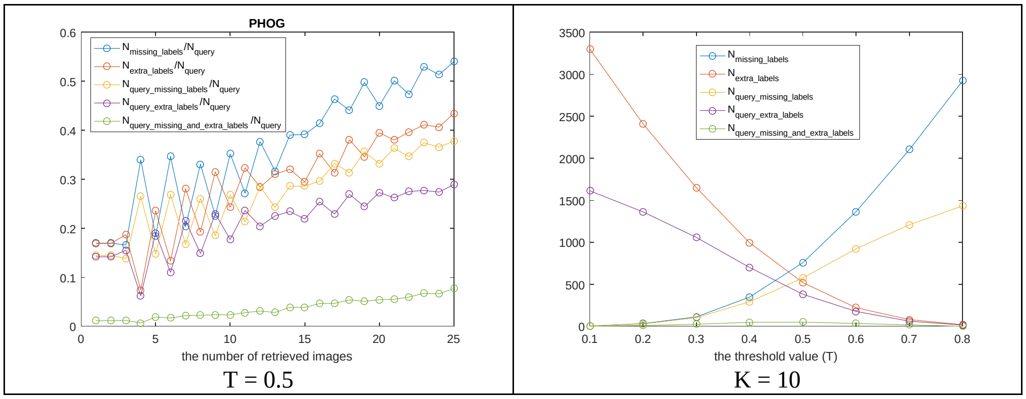


Figure 3. Experiment 1: Annotation and retrieval result of one example abdominal image



T = 0.5

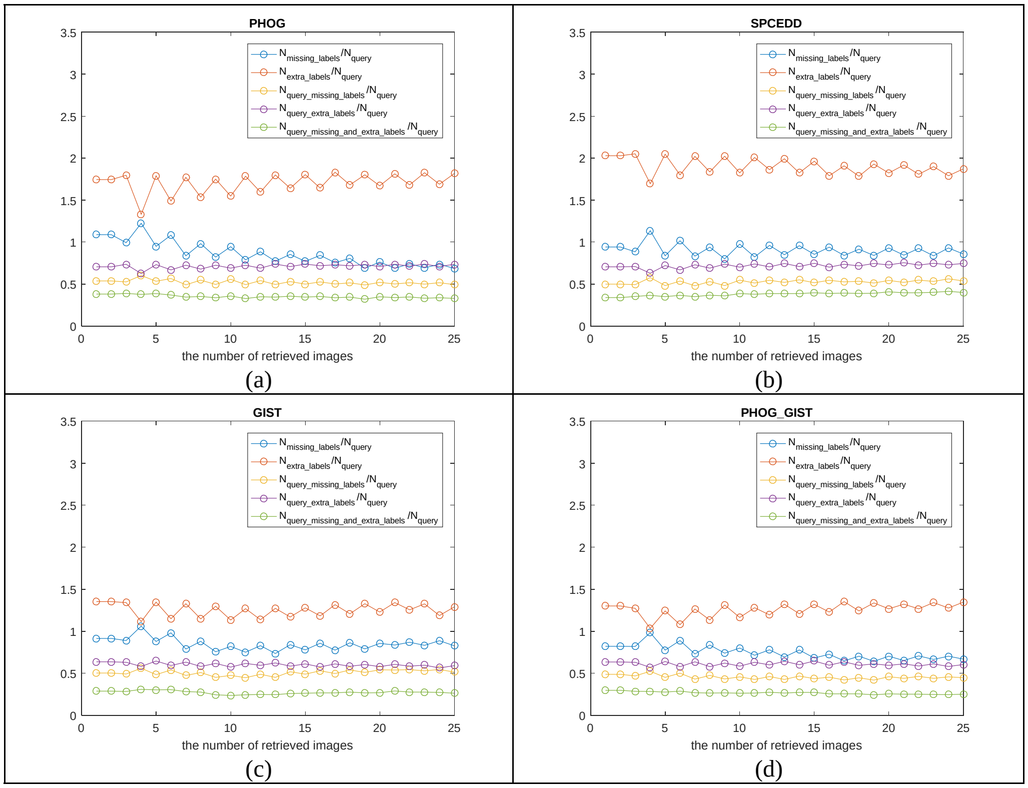
K = 10

Figure 4. Experiment 1: annotation performance with respect to K or T (abdominal set)

3.2 Experiment 2

In this experiment, we separate images in the labeled abdominal dataset into two sets. One contains all the slices generated from the first 5 patients' CT volumetric data. We use all the slices in this set as query images. The other contains all the slices obtained from the other 25 patients' CT volumetric data. All the images in this set are used to annotate the query images. Thus, the images in the query database and the images in the retrieval database are from different patients. As a result, the visual differences between the images in the query database and the images in the retrieval database are larger than that in the Experiment 1. There are 427 images in the query database and 1723 images in the retrieval database for the abdomen data. Figure 5(a) shows the annotation performance using PHOG with respect to the number of retrieved images *K*, when the threshold value *T* is set to be 50%. Compared to the corresponding graphs in Experiment 1, the results are notably worse, which is expected. We may account for the differences between the content of two slice images with two factors: one is contributed by the inter-patient difference at similar body locations and the other is contributed by the inter-patient difference at different body locations. Ideally, the system will be tolerant of the former difference, but discriminative on the latter. However, we found this is very challenging to achieve, because the inter-patient differences at similar body locations are quite large. For example, Figure 6(a) shows the examples of the slice of different patients in which the right kidney (yellow) appears for the first/second time in the body slice stack (from top to bottom) (the second slice is shown instead if the right kidney is too small in the first slice). Among these example slices which are approximately located at the same body axial location on different patients, the shape, size, and location of each organ, the number of organs, the shape and size of the body, the shape, size and location of the non-labeled regions such as bones, are all quite different. The same observation applies to the images in Figure 6(b), and (c).

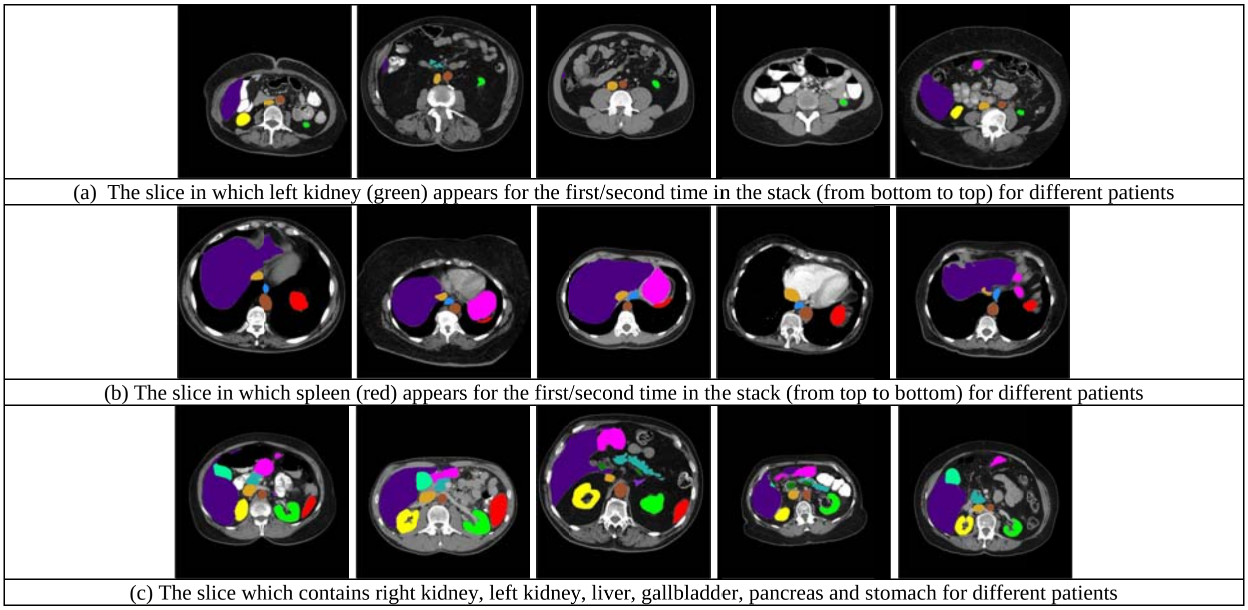
We have considered several approaches to improve the retrieval performance. For either classification or nearest neighbor searching, feature extraction is an important step. Therefore one of our efforts is to explore alternative feature descriptors to see if they can achieve better performance than PHOG. SPCEDD and GIST are among the features we have tested. Our experiments show that for the abdomen data, the result of using GIST is better than that of PHOG, and the result of PHOG is better than that of SPCEDD. For the cervix data, the result of PHOG is better than that of GIST, and the result of GIST is better than that of SPCEDD. Therefore, we also test the combined feature of PHOG and GIST. In addition to examining additional features, we also evaluate PHOG performance using different parameter values (number of bins *B* and number of levels *L*). Results were that, the performance of *B* = 8 is similar to *B* = 16; the performance of *L* = 3 is similar to *L* = 2, and both are better than the performance of *L* = 1. Roughly speaking, the images in each row in Figure 6 look more similar to each other than to others in other rows (which are relatively more distant locations along the body z-direction). This observation motivates the idea of the coarse-to-fine, two-step approach described in Section 2.6. To train the supervised classifier, we experimentally divide the series of slices into different body segment groups using heuristic rules. For example, for abdomen data, four groups are generated. Specifically, given the slices ordered from bottom to top along the z-direction for each patient, group 2 contains all the slices that include both right kidney and left kidney; group 1 contains all the slices that are below the first slice in group 2; group 4 contains all the slices that are above the last slice in which the liver is included; and group 3 contains all of the remaining slices. Only the images in the retrieval database are used to train the classifier. The features used by the classifier are PHOG and GIST. We use the SMO (Sequential Minimal Optimization) implemented in Weka [13] to do the classification. The annotation performance of the coarse-to-fine approach for abdomen data is shown in Figure 7, and shows performance improvement as compared to Figure 5. Figure 8 shows the retrieval results of a query image that demonstrates the improvement by using the above approach. The first row is the query image. The results of three approaches: PHOG with NN, PHOG+GIST with NN, and PHOG+GIST with coarse-to-fine are shown in the second, third, and fourth row respectively.



(a) (b)

(c) (d)

Figure 5. Experiment 2: annotation performance with respect to K for abdomen data (T = 50%)



(a) The slice in which left kidney (green) appears for the first/second time in the stack (from bottom to top) for different patients

(b) The slice in which spleen (red) appears for the first/second time in the stack (from top to bottom) for different patients

(c) The slice which contains right kidney, left kidney, liver, gallbladder, pancreas and stomach for different patients

Figure 6. Inter-patient differences

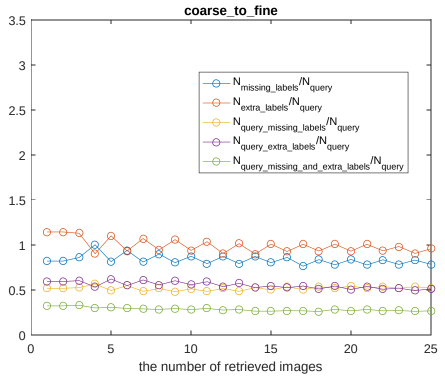


Figure 7. Performance of the coarse-to-fine approach for the abdomen data

3.3 Experiment 3

This experiment aims to test the method for annotating figure query images using all the labeled slices extracted from the volumetric data. It serves to provide discussion and gain more insights for our goal – annotating figures in biomedical articles, which will be used for VQA. Since we do not have ground truth labels for the figures, we cannot quantitatively evaluate the results, so we use a qualitative, “proxy” evaluation method, based on how well our method performs in retrieving images similar to the query image. Because good retrieval performance indicates good annotation performance, we visually examine the retrieved results on a limited number of query figure images in order to get preliminary idea on the effectiveness and challenges of adapting this method to a figure query dataset. The figure set that we tested contains about 100 abdomen figures downloaded from Open-i. The retrieval dataset contains all of the labeled abdomen slices. Again, our evaluation is preliminary and qualitative: among the results, some appear to be satisfactory while others do not. The unsatisfactory results may be due to several reasons: 1) the large visual differences between the figures in Open-i and the slices used in the retrieval database; 2) the feature descriptor is not effective enough; 3) all the regions in the image are treated basically similarly (although regions having high gradients are weighted higher in PHOG calculation); 4) the figures are very diverse with respect to image content and image quality/resolution. Therefore, our future work includes collecting more labeled data, objective and subjective evaluation of the retrieval results, and testing more features and distance measures.

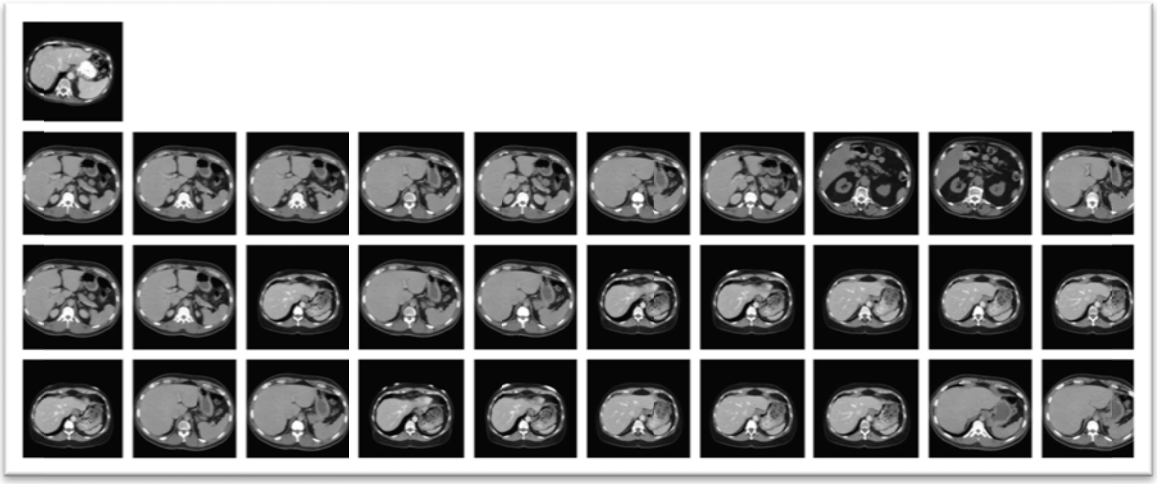


Figure 8. Retrieval results with different approaches

4. CONCLUSIONS

In this paper, we aim to automatically annotate a 2-D axial view CT image with multiple anatomical organ terms. This multi-label image annotation method is based on using content-based image retrieval techniques. We work on a set of labeled multi-organ 2-D CT slices generated from the CT volumetric dataset provided by the MICCAI 2015 organ segmentation challenge. We extract the features PHOG, SPCEDD, and GIST for finding similar images. We adapt a weighted voting scheme to assign multiple labels to the query image by combining the labels of the similar images identified by the method. We propose a coarse-to-fine retrieval strategy which integrates the classification at the coarse level with nearest-neighborhood search at the fine level. We carry out two experiments using the MICCAI 2-D CT datasets to quantitatively compare and evaluate features, parameters and retrieval approaches. We also test our method for annotating Open-i figure query images and discuss the results.

ACKNOWLEDGEMENT

This research was supported 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] Shin, H., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R. M., “Learning to read chest X-rays: recurrent neural feedback model for automated image annotation”, IEEE CVPR, (2016).

[2] Rahman, M.M., Antani, S.K., Demner-Fushman, D., Thoma, G.R., “Biomedical image representation approach using visualness and spatial information in a concept feature space for interactive region-of-interest-based retrieval”, J Med Imaging, 2(4), (2015).

[3] Kumar, A., Dyer, S., Kim, J., Li, C., Leong, P.H.W., Fulham, M., Feng, D., “Adapting content-based image retrieval techniques for the semantic annotation of medical images”, Computerized Medical Imaging and Graphics, 49, 37-45 (2016).

[4] Li, X., Chen, L., Zhang, L., Lin, F., Ma, W.Y., “Image annotation by large-scale content-based image retrieval”, Proceedings of the 14th ACM International Conference on Multimedia, 607-610 (2006).

[5] Noel, G.E., Peterson, G.L., “Context-driven image annotation using ImageNet”, Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference, 462-467 (2013).

[6] https://www.synapse.org/#!Synapse:syn3193805/wiki/

[7] http://www.radiantviewer.com/dicom-viewer-manual/change_brightness_contrast.htm

[8] Liu, Y., Zhang, D., Lu, G., Ma, W. “A survey of content-based image retrieval with high-level semantics”, Pattern Recognition, 40(1), 262-282 (2007).

[9] Bosch, A., Zisserman, A., Muñoz, X., “Representing shape with a spatial pyramid kernel,” International Conference on Image and Video Retrieval, Amsterdam, The Netherlands (2007).

[10] http://www.lire-project.net/

[11] Chatzichristofis, S.A., Boutalis, Y.S., “CEDD: color and edge directivity descriptor – a compact descriptor for image indexing and retrieval.”, 6th International Conference in advanced research on Computer Vision Systems (ICVS), Lecture Notes in Computer Science (LNCS), 312-322 (2008).

[12] Oliva, A., Torralba, A., “Modeling the shape of the scene: a holistic representation of the spatial envelope”, International Journal of Computer Vision, 42(3), 145-175 (2001).

[13] http://www.cs.waikato.ac.nz/ml/weka/