Towards the Creation of a Visual Ontology of Biomedical Imaging Entities

Matthew S. Simpson, PhD; Daekeun You, PhD; Md Mahmudur Rahman, PhD; Sameer K. Antani, PhD; George R. Thoma, PhD; and Dina Demner-Fushman, MD, PhD

Lister Hill National Center for Biomedical Communications
U. S. National Library of Medicine, Bethesda, MD

ABSTRACT

Image content is frequently the target of biomedical information extraction systems. However, the meaning of this content cannot be easily understood without some associated text. In order to improve the integration of textual and visual information, we are developing a visual ontology for biomedical image retrieval. Our visual ontology maps the appearance of image regions to concepts in an existing textual ontology, thereby inheriting relationships among the visual entities. Such a resource creates a bridge between the visual characteristics of important image regions and their semantic interpretation. We automatically populate our visual ontology by pairing image regions with their associated descriptions. To demonstrate the usefulness of this resource, we have developed a classification method that automatically labels image regions with appropriate concepts based solely on their appearance. Our results for thoracic imaging terms show that our methods are promising first steps towards the creation of a biomedical visual ontology.

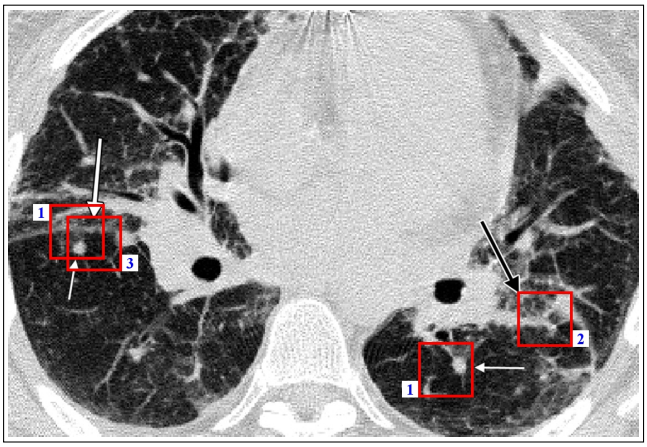
INTRODUCTION

Information exists in a variety of forms. Common sources of information in the biomedical domain include the images and text found within scientific articles, patient health records, and specialized medical imaging databases. Given the rapid pace of scientific discovery and the continued growth of the health care industry, relevant images and text are often buried among volumes of irrelevant data, making it difficult for researchers and clinicians to quickly and reliably assimilate the most relevant information for a particular need. Not surprisingly, information extraction and retrieval are essential tasks required for achieving many of the ultimate goals of biomedical informatics research and development.[1] These goals include supporting clinical decisions, creating rich document summaries, and answering clinical questions.

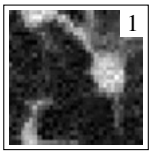
Unfortunately, biomedical images cannot be easily understood when they are removed from their original context.[2] Consider the task of retrieving images from a collection of biomedical articles. It is common for authors to include in their reports images that help explain or demonstrate some concept introduced within the full text of their articles. For example, an author might report on the size and shape of some nodules depicted in a patient's lung CT scan, using the visual characteristics of these regions as a basis for a possible cancer diagnosis. If an image retrieval system separates this image from its associated text, as is often done by content-based image retrieval systems, it disregards the meaning attached to the appearance of the nodules. Moreover, automatically contextualizing the important regions within an image having no associated text is an impossible task because the appearance of such regions cannot be related in a meaningful way to other entities. Whereas ontological resources exist for relating the meaning of textual entities, no such resource exists for relating the appearance of visual entities.

In this paper, we propose the creation of a visual ontology of biomedical imaging entities, and we report on our initial progress towards reaching this goal. Such a resource serves two primary functions. First, it defines a set of visual entities and maps the appearance of these entities to their textual descriptions. This association creates a bridge between the visual characteristics of important regions within an image and their semantic interpretation. For example, an area of a CT scan having a slightly bright and hazy appearance (a visual entity) might be mapped to "ground-glass opacity" (its textual description). Second, a visual ontology defines relationships among the entities. While not a requirement, if the ontology maps visual entities to concepts within existing textual ontologies, the relationships among visual entities can be aligned with those of their corresponding textual entities.

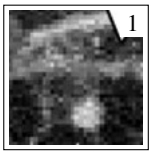
A comprehensive resource performing these functions would have a variety of practical applications, especially in the area of image retrieval. Text-based image retrieval systems searching for images within a collection of biomedical articles commonly represent and retrieve them according to their associated captions. Aided by a visual ontology, such a system might extract concepts from the text of a query, map these concepts to their visual characteristics, and then use



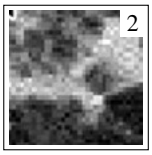
Caption: Figure 18a. Recurrence of sarcoidosis in a patient with bilateral lung transplants for end-stage pulmonary fibrosis secondary to sarcoidosis. (a) High-resolution CT image, obtained more than 16 weeks after lung transplantation, shows multiple pulmonary nodules (small white arrows), nodularity along the right major fissure (large white arrow), peribronchial thickening, ground-glass opacities, and patchy architectural distortion (black arrow). (b) High-power photomicrograph (original magnification, ×200; H–E stain) of a specimen from bronchoscopic biopsy reveals multiple discrete nonnecrotizing granulomas (arrows) in the wall of a bronchiole (*).



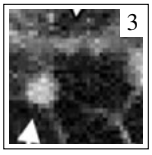
Marker: Arrow
Number: Plural
Color: White
Size: Small
Description: Multiple pulmonary nodules
RadLex ID: RID3877



Marker: Arrow
Number: Plural
Color: White
Size: Small
Description: Multiple pulmonary nodules
RadLex ID: RID3877



Marker: Arrow
Number: Singular
Color: Black
Size: –
Description: Patchy architectural distortion
RadLex ID: RID34261



Marker: Arrow
Number: Singular
Color: White
Size: Large
Description: The right major fissure
RadLex ID: RID1374

(a) Original image

(b) Regions of interest

Figure 1: An example figure (a) from "Postoperative Complications of Lung Transplantation: Radiologic Findings along a Time Continuum" by Krishnam et al. [3] is shown with its visual and textual regions of interest (b). Each region pairs a patch taken from the original image with a description of the patch taken from the image's caption.

this information to search the content of images as well as their captions. Alternatively, a content-based image retrieval system might extract important regions from an example query image, map the visual characteristics of these regions to textual concepts, and then use these concepts to search image captions.

For our initial feasibility evaluation, we limit our efforts at constructing a visual ontology to thoracic CT scans and their associated captions taken from a collection of biomedical articles. These images commonly contain regions of interest (ROIs) that are denoted visually by overlain markers such as arrows or asterisk symbols. In addition, the captions associated with these images often contain descriptions of the ROIs and reference them using the visual attributes of their markers. For example, a caption might describe a region marked by a "large white arrow." Thus, the bridge between a region's visual characteristics and its textual description is frequently made explicit through the mention of its marker. Because textual descriptions of many radiologic imaging observations exist in RadLex,[4] we can populate our visual ontology with these ROIs and infer relationships among them using their descriptions.

In Figure 1, which we will further discuss when describing our methods, we show a thoracic CT scan and its associated ROIs. We refer to the boxed regions within the image as visual ROIs and their descriptions and marker attributes as textual ROIs. By mapping ROI descriptions to RadLex concepts (denoted in Figure 1b by their identifiers), we align the appearance of image regions with semantic concepts and inherit from RadLex their ontological relationships.

Basing our visual ontology on the recognition of ROIs and their descriptions, the goals of our current work are (i) to automatically extract and pair the visual and textual ROIs contained within images and their captions and (ii) to use this resource to automatically label visual ROIs with appropriate textual descriptions based solely on their appearance.

In order to achieve these goals, we developed ROI extraction and classification methods and evaluated these approaches on a gold standard of manually annotated ROIs that we created. Our extraction methods utilize a combination of statistical and rule-based natural language and image processing techniques to label local image regions with their textual descriptions. Our classification method then uses these labeled regions to train a classifier for labeling visual ROIs with one of five thoracic imaging concepts. Thus, our methods are capable of automatically mapping the appearance of visual entities within images to a limited set of concepts. Our experimental results show that our methods are promising first steps towards the creation of a visual ontology of biomedical imaging entities.

BACKGROUND

The images contained in biomedical articles are not always self-explanatory, and much of the information required for their comprehension can be found in the text of the articles in which they appear.[5] Therefore, methods for integrating text with image content has been the subject of much research.

The medical retrieval track of the ImageCLEF[6] evaluations has been an important catalyst for advancing methods of integrating textual and visual information within the biomedical domain. Participants of these evaluations are provided several multimodal topics, and they are tasked with retrieving for each topic the most relevant images from a collection of biomedical articles. Although an exhaustive account of the retrieval methods implemented as part of these evaluations is not feasible, Müller et al.'s retrospective[7] is an appropriate starting point.

The best-performing systems at the ImageCLEF evaluations have historically relied upon traditional text-based information retrieval methods. However, recent systems have shown encouraging progress towards combining these methods with approaches from content-based image retrieval. One such approach is the use of VisMed terms[8] for image indexing and retrieval. Our biomedical visual ontology is aimed at continuing this progress by combining text-based and content-based image features in semantically meaningful ways.

Outside the biomedical domain, methods of integrating text with image content have been applied to a variety of computer vision tasks. Though too numerous to discuss in detail, some of these tasks include video retrieval,[9,10] the retrieval of art images,[11] the annotation of animals in photographs,[12] the labeling of faces in newspaper photographs,[13] and the generation of natural language descriptions of scenes.[14]

METHODS

We populate our visual ontology by extracting regions of interest (ROIs) from images contained in biomedical articles and pairing these entities with textual descriptions taken from their associated captions. To demonstrate the utility of a visual ontology, we then use these labeled regions to train a classifier for automatically assigning concepts to ROIs having no associated text. However, in order to evaluate our ROI extraction and classification methods, we require a set of manually annotated ROIs to serve as our gold standard. We describe our effort at manually annotating ROIs first.

Manual Annotation of Regions of Interest

We chose to limit our annotation effort to thoracic CT scans. Such images exhibit high regularity and account for a large portion of the images publicly available as part of the 2010 ImageCLEF medical retrieval track data set.[15]

To obtain a set of images and associated captions to annotate, we searched the ImageCLEF collection using each concept contained in the article entitled "Fleishner Society: Glossary of Terms for Thoracic Imaging" by Hansell et al.[16] We used this glossary of terms as an aid for retrieving thoracic images in part because it is a source for RadLex. Thus, if we populate our visual ontology with ROIs whose descriptions are terms in the glossary, we can use RadLex to define the relations among the visual entities. We used the Essie[17] information retrieval system, which expands query terms along synonymy relationships in the Unified Medical Language System®(UMLS®),[18] to index the ImageCLEF captions and to retrieve the most relevant images for each term in the glossary. However, because the number of images we retrieved using this approach was prohibitively large for our manual annotation effort, we then narrowed the set of images by discarding those that were not retrieved by one of a few frequently occurring terms. These terms, which are in the "imaging observation" (RID5) branch of the RadLex tree, included "ground-glass opacity" (RID28531), "honeycombing" (RID35280), and "tree-in-bud pattern" (RID35654), among others.

Having retrieved a subset of the thoracic images and their associated captions from the 2010 ImageCLEF medical retrieval track data set, we then manually annotated their textual and visual ROIs. One author (Simpson) annotated the textual ROIs in the retrieved images' captions, and one author (You) annotated the images' visual ROIs and paired them with their corresponding annotated textual ROIs.

For the manual annotation of textual ROIs, we identified markers within an image caption, their visual attributes, and descriptions of their marked regions. Following the concept annotation guidelines from the 2010 i2b2/VA Challenge,[19] we developed the rules described below for annotating these elements.

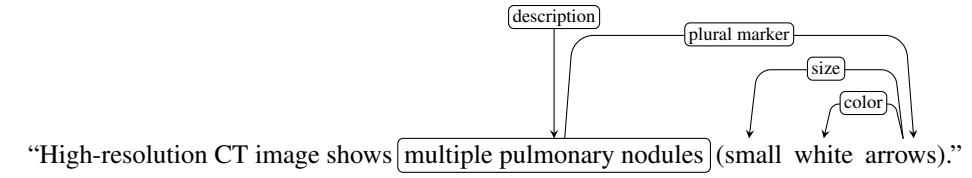


Figure 2: An example caption sentence is shown with its annotated textual region of interest. Annotations are depicted as a dependency graph having the region description as its root.

1. We only annotated single-word nouns, symbols, and letters as ROI markers, and we identified them as being either singular or plural. We assessed a marker's plurality by inspecting a caption's associated image because authors often fail to unambiguously identify the number of visible markers. Moreover, some markers, such as the asterisk symbol, cannot be made plural, yet they are often placed in multiple locations within an image. Example ROI markers include the words "arrow," "arrowhead," and their plurals.
2. We only annotated single-word adjectives describing color, shape, and size as ROI marker attributes. Example ROI marker attributes include the words "white," "open," and "large."
3. We only annotated complete noun, adjective, and verb phrases as ROI descriptions.
   (a) We included at most one prepositional phrase as part of a ROI description if it could be eliminated by rearranging the description. For example, phrases such as "a patchy ground-glass opacity area" and "a patchy area of ground-glass opacity" were both annotated as descriptions.
   (b) We included conjoined phrasal elements as part of a ROI description if the elements shared a common set of modifiers. For example, we annotated "patchy ground-glass and airspace opacities" as a single description.
   (c) We included non-restrictive appositive phrases that could otherwise be annotated as ROI descriptions as part of broader descriptions containing the phrases they define. For example, we annotated "branching nodular structures (tree-in-bud pattern)" as a single description.

Figure 2 shows a fragment of the image caption depicted in Figure 1 with the aforementioned textual ROI annotations represented as a dependency graph. The root of the graph is the description, which for this sentence is the phrase "multiple pulmonary nodules." The nodular regions are denoted visually by multiple "arrow" markers, which can be distinguished from the other markers appearing in the image by their "small" size and "white" color.

For the manual annotation of visual ROIs, we identified within images local patches denoted by visible markers. We developed the following additional guideline for annotating visual ROIs.

4. We only annotated local patches within an image using non-rotated one hundred pixel by one hundred pixel square boxes overlaid on the image at its full resolution.
   (a) For patches denoted by directional markers, we placed the annotation box in the direction of the marker. For example, if a visual ROI was marked by an arrow, we placed the annotation box such that the arrow was pointing to the box's center, and the tip of the arrow was touching the box's boundary.
   (b) For patches denoted by non-directional markers, we centered the annotation box on the marker. For example, if a visual ROI was marked by an asterisk symbol, we placed the annotation box such that the asterisk was in the center of the box.

Having annotated an image's textual and visual ROIs using the aforementioned guidelines, we then extracted from the image the local patches enclosed within the annotation boxes and paired them with the ROI descriptions annotated in the image's caption. We utilized the ROI marker attributes for selecting the corresponding visual ROIs for a given textual ROI. Recall that because a textual marker may be plural, a single description may correspond to multiple patches.

The textual and visual ROIs previously introduced in Figure 1 are a result of our manual annotation effort. Figure 1a depicts the original image and its annotated visual ROIs. Note that the image's caption, in addition to containing a description of this image (figure 18a), also includes a description of another image (figure 18b), indicating that this image is part of a multi-panel figure. We ignored the portions of a caption that refer to other images within a multi-panel figure when performing our annotation. Figure 1b shows the image's paired textual and visual ROIs. It depicts the extracted image patches with their corresponding descriptions, markers, and marker attributes. Because the "arrow"

marker that is "small" and "white" is plural, we used the two such arrows to associate "multiple pulmonary nodules" with their visual characteristics, as shown in the first two patches of Figure 1b. Similarly, we associated the appearance of the patch pointed to by an "arrow" marker that is "black" with "patchy architectural distortion," and we associated the appearance of the patch marked by an "arrow" that is "large" and "white" with "the right major fissure."

Having created a gold standard set of manually annotated ROIs, we then evaluated our automatic textual and visual ROI extraction methods as well as trained our visual ROI classifier. We describe these methods next.

Automatic Textual Region of Interest Extraction

Our automatic textual ROI extraction method combines rule-based and statistical natural language processing techniques to identify textual markers, marker attributes, and descriptions within image captions. First, we preprocess the captions of multi-panel figures following Demner-Fushman et al. [2] in order to distinguish text referring to a specific image. This segmentation is necessary for some captions, like the one shown in Figure 1a, to avoid falsely identifying textual ROIs for absent images. The caption segmentation method uses regular expressions to locate sub-caption labels such as "(a)" or "(a–c)," and it applies rules for attributing sub-caption text to each of the resulting labels. Next, we use a maximum entropy sentence detector to split the image's caption into sentences. Finally, we identify textual ROIs in each resulting sentence following Apostolova and Demner-Fushman.[20] The textual ROI extraction method first identifies ROI markers and their attributes using regular expressions. It then produces a full syntactic parse of the sentence using Stanford University's unlexicalized probabilistic context-free grammar parser.[21] The method selects the noun phrase nearest a given ROI marker as its associated description.

Automatic Visual Region of Interest Extraction

Similar to our textual ROI extraction approach, our automatic visual ROI extraction method utilizes a combination of rule-based and statistical image processing techniques to identify and group the visible markers within images. First, we locate arrows, arrowheads, and asterisk symbols using the marker recognition algorithm of You et al.[22] This method utilizes Markov random fields and hidden Markov models to identify a marker's location within an image, its shape, color, size, and the direction it is pointing. Next, we use the textual marker attributes identified in an image's caption to pair the recognized visible markers with their textual mentions. A marker's absolute visible attributes, such as its shape and color, are determined by the aforementioned marker recognition algorithm, and these are easily paired with textual attributes. However, relative marker attributes, such as size, are more challenging. For this purpose, we group all markers of a given shape and color by their size using a dynamic time warping algorithm. We can then distinguish the relative size of the markers in the resulting groups as well as the number of markers sharing the same attributes. Finally, having paired the visible markers with their textual mentions, we extract from the image a local region for each marker and name the patches according to the descriptions of their corresponding textual ROIs. Although precise region boundaries may be delineated using a variety of segmentation techniques,[23] we extract one hundred pixel by one hundred pixel patches, orienting them relative to their markers according to our annotation guidelines.

Automatic Visual Region of Interest Classification

Our automatic visual ROI classification method uses a supervised classifier trained on a set of paired textual and visual ROIs in order to label visual ROIs having no associated text with appropriate concepts. Recall that for constructing our gold standard set of manually annotated ROIs, we consulted a glossary of thoracic imaging terms to retrieve a set of CT scans. While these concepts must occur somewhere in the retrieved images' captions, because they do not necessarily correspond to any annotated textual ROIs, we cannot use them as our class labels. Therefore, we mapped the annotated ROI descriptions to their corresponding concepts in the glossary and selected as our class labels five terms for which we had an adequate number of paired ROIs to use for training. These terms – several of which can be found in RadLex – include "bronchiole" (RID1298), "consolidation," "cyst" (RID3890), "ground-glass opacity" (RID28531), and "mosaic attenuation pattern." As bronchioles are not normally visible in thoracic CT scans,[16] our "bronchiole" label is actually a metaclass containing abnormal bronchial pathologies such as "bronchial wall thickening" and "traction bronchiectasis."

Like other lung tissue classification approaches,[29,30] the performance of our method depends greatly on our underlying representation of visual ROI content. For our current work, we represent the content of each visual ROI as a 487-dimensional vector, which combines the 8 texture-related features shown in Table 1. While a discussion of the strengths and weaknesses of these eight features is beyond the scope of this paper, we recognize that no single representation is adequate for describing the content of all of our visual ROIs. Therefore, we assume these features to be complementary.

Table 1: Texture-related features for representing the content of visual ROIs.

Feature	Dimensionality
Image moments*	3
Gray-level co-occurrence matrix moments† [24]	20
Autocorrelation coefficients	25
Edge frequency	25
Gabor filter descriptor‡	60
Tamura descriptor‡ [25]	18
Color and edge directivity descriptor‡ [26]	144
Fuzzy color and texture histogram‡ [27]	192
Combined texture feature	487

*Moments include mean, standard deviation, and skewness.
†Moments are computed on four image orientations and include energy, maximum probability, entropy, contrast, and inverse difference.
‡Feature computed using the Lucene Image Retrieval library. [28]

Having a representation for the content of visual ROIs and a set of five thoracic imaging concepts with which to label them, we can train our visual ROI classifier. In our current work, we evaluate the accuracy of three classifiers including a multi-class support vector machine (SVM), an artificial neural network (ANN), and a multinomial naïve Bayes (NB) classifier. We report on the evaluation of our ROI extraction and classification methods in the remaining sections.

EVALUATION

We evaluated our ROI extraction and classification methods against the gold standard we created through our manual annotation effort. For our extraction methods, we measured the precision and recall of correctly identifying ROI elements, and for our visual ROI classification method, we measured the classifiers' ten-fold cross-validation accuracy. Given that our current work only represents our first steps toward the creation of a visual ontology, we evaluated our ROI extraction and classification methods under varying degrees of strictness, which we describe next.

For our textual ROI extraction method, we considered several strategies for measuring the correctness of the phrases our method extracted as ROI descriptions. Similar to the notions of correctness proposed by Olsson et al., [31] we evaluated the success of our ROI description extraction approach according to the following three criteria.

Exact Both the left and right boundaries of an extracted ROI description are required to exactly match those of the manual annotation.

Inexact Only one boundary of an extracted ROI description is required to match that of the manual annotation. Thus, the extracted description either begins or ends at the correct position.

Overlapping Neither boundary of an extracted ROI description is required to match that of the manual annotation, but the extracted and annotated regions must overlap.

Our visual ROI extraction and classification methods are sensitive to errors produced by our textual ROI extraction method. For example, our visual ROI extraction method uses the knowledge of textual ROIs in an image's caption to help identify and pair visible markers with their corresponding textual markers. In addition, because our visual ROI classification method uses the resulting paired ROIs as labeled training examples, it is sensitive to errors made during the ROI pairing process. In order to account for and better understand this cascading of errors, we evaluated the success of our visual ROI extraction and classification methods according to the following two scenarios.

Ideal The input and training data are derived from our manual annotation effort. Thus, our visual ROI extraction method pairs extracted visual ROIs with manually annotated textual ROIs, and we train our visual ROI classifier with manually annotated and paired ROIs.

Actual The input and training data are derived from our extraction methods. Thus, our visual ROI extraction method pairs extracted visual ROIs with extracted textual ROIs, and we train our visual ROI classifier on this extracted set of automatically paired ROIs.

Our visual ROI classification method depends on texture-related features that we extract from the local patches associated with each ROI. However, we did not directly evaluate the utility of our current approach for sizing, placing, and orienting

Table 2: Manual annotation of ROIs.

Data	Total
Images	298
ROIs	1052
"bronchiole"	178
"consolidation"	57
"cyst"	51
"ground-glass opacity"	176
"mosaic attenuation pattern"	15

Table 3: Automatic textual ROI extraction.

Extracted Elements	Precision	Recall	F_1 Score
Marker	0.9619	0.8688	0.9130
Marker, attributes	0.9238	0.8344	0.8768
Marker, attributes, description (overlapping)	0.8000	0.7226	0.7593
Marker, attributes, description (inexact)	0.7881	0.7118	0.7480
Marker, attributes, description (exact)	0.5429	0.4903	0.5153

Table 4: Automatic visual ROI extraction.

Extracted Elements	Precision	Recall	F_1 Score
Paired ROIs (ideal)	0.7939	0.6902	0.7384
Paired ROIs (actual)	0.3847	0.2784	0.3230
Markers	0.1675	0.7922	0.2766

Table 5: Automatic visual ROI classification.

ROI Classifier	Accuracy (%)
SVM (ideal, filtered)	77.44
ANN (ideal, filtered)	76.15
ANN (ideal, unfiltered)	71.91
SVM (ideal, unfiltered)	70.65
NB (ideal, filtered)	66.92
SVM (actual, unfiltered)	61.29
NB (ideal, unfiltered)	56.39
ANN (actual, unfiltered)	54.84
NB (actual unfiltered)	45.48

these patches. Recall that our visual ROI extraction method simply represents regions as uniformly sized boxes placed according to our annotation guidelines. Therefore, in order to account for errors associated with our region placement strategy, we evaluated the success of our visual ROI classifiers on the following two sets of training examples.

Unfiltered The training examples are an unmodified set of paired ROIs produced by our ROI extraction methods. Note that this set can either be an ideal or actual set of labeled image patches corresponding to the five concepts we selected as our class labels.

Filtered The training examples are the set of paired ROIs produced by our ROI extraction methods from which we remove ROIs whose appearance is obscured by our current region placement strategy. For example, we remove ROIs if their associated patches are not sized or placed in a way that adequately reflects the regions' descriptions. We also remove ROIs whose patches are centered on non-directional markers because our current image processing methods do not account for such markers.

RESULTS

Table 2 shows results related to our manual annotation effort. Using a small set of concepts defined in a glossary of thoracic imaging terms, we retrieved a total of 298 CT scans, similar to the one shown in Figure 1, from the 2010 ImageCLEF medical retrieval track data set. For the retrieved images and their associated captions, we manually annotated and paired a total of 1052 ROIs. After mapping the annotated ROI descriptions back to terms in the glossary (the terms used to retrieve the images were not necessarily annotated as ROI descriptions), we obtained a total of 477 paired ROIs relating to 5 glossary terms. We then used the ROIs associated with these concepts to train our visual ROI classifier. Table 2 provides the number of training examples for each concept.

Table 3 shows results for our textual ROI extraction approach. It gives our method's precision, recall, and F_1 score for correctly identifying the textual ROI elements that we manually annotated in our set of 298 image captions. For automatically extracting only ROI markers, our method achieves an F_1 of 0.91. However, when we require that these markers be correctly identified with all of their visual attributes, this score is reduced to 0.87. Similarly, our method achieves an F_1 of 0.52 for exactly identifying ROI descriptions with their correct markers and attributes, but this score improves to 0.76 as the description boundary conditions are relaxed.

We give the precision, recall, and F_1 score of our visual ROI extraction approach in Table 4. For identifying the visible markers in our set of 298 images, our method achieves an F_1 of 0.28, a low score due to the extraction of many false

positives. However, given knowledge of the textual ROIs in the images' captions, this noise is greatly reduced. When using our gold standard set of annotated textual ROIs as input, our method achieves an F_1 of 0.74 for pairing identified markers with their corresponding textual ROIs. When using the actual textual ROIs identified by our extraction method, the combined errors associated with these approaches reduces this score to 0.32.

Finally, Table 5 shows the ten-fold cross-validation accuracy of our visual ROI classifiers. When trained on the unfiltered gold standard set of ROIs, the ANN classifier outperformed the SVM and NB classifiers, achieving a cross-validation accuracy of 71.91. After removing from our ideal training set the ROIs whose appearance is negatively impacted by our current region placement and segmentation strategy, the set of 477 ROIs corresponding to our 5 classes is reduced to 390 examples. The SVM classifier (polynomial kernel, $C = 1.0$, $\epsilon = 1.3$) performed best on this reduced set of examples, achieving an accuracy of 77.44. Similarly, when trained on the unfiltered set of paired ROIs produced by our extraction methods, the SVM classifier (polynomial kernel, $C = 1.0$, $\epsilon = 1.0$) also performed best, obtaining an accuracy of 61.29. Because many of the automatically paired ROIs are incorrectly labeled, the number of training examples in this set is further reduced to 310 ROIs. For choosing our SVM kernel and parameters, we experimented with using a radial basis function, but the polynomial kernel consistently resulted in a better cross-validation accuracy on our various data sets.

DISCUSSION

Our results show that our ROI extraction and classification methods are promising first steps towards the creation of a visual ontology of biomedical imaging entities. Having evaluated our methods against a set of manually annotated ROIs, we gained insight into the strengths and weaknesses of our current extraction and classification methods.

Our textual ROI extraction method is successful at identifying markers within image captions. The method's success provides evidence that our caption segmentation approach is adequate for attributing sub-caption text to images within multi-panel figures and that the regular expressions we use for identifying markers capture those commonly seen in thoracic CT scans. However, our method has trouble identifying single-letter markers (they can easily be confused as sub-caption labels), and the plurality of single-letter and symbolic markers. A marker's plurality is often not possible to discern textually. For example, authors commonly mark more than one region of an image with an asterisk symbol, but because an asterisk is not a regular word, it can be cumbersome to indicate the presence of more than one.

Unfortunately, our method has difficulty identifying descriptions corresponding to the textual markers. We can attribute our method's tendency to incorrectly identify the boundaries of ROI descriptions in part to the difficulties associated with parsing biomedical text. However, given a correct parse, our results suggest that selecting the nearest noun phrase to a marker as its corresponding description is not always adequate. Consider, for example, the following sentence.

> Transverse high-resolution CT scan obtained in a 46-year-old man shows a combination of grades 1 (short arrow), 2 (long arrow), and 3 (arrowheads) bronchial wall thickening in the left lower lobe.[32]

The ROI description corresponding to the "long arrow" in this sentence would ideally be "grade 2 bronchial wall thickening," but the wording of the sentence makes it difficult to reconstruct such a description simply by analyzing nearby noun phrases. For such captions, the consideration of syntactic dependencies may aid in improving our method for identifying ROI descriptions and their correct boundaries.

Our visual ROI extraction method produces many false positives when identifying the visible markers within images. The low precision of our approach results from the numerous areas of high contrast commonly seen in CT scans of lung tissue, which our algorithm sometimes confuses as arrowheads. Though the number of false positives is greatly reduced when our algorithm is made aware of the identified textual markers and their attributes, preprocessing the images to reduce noise may aid in improving our visual ROI extraction method.

However, even with the ability to precisely identify all markers within an image, automatically extracting meaningful regions corresponding to them is not a straightforward task. In our current work, we simply extract one hundred pixel by one hundred pixel square patches for each marker, but the size and placement of these patches do not always adequately represent the regions described in the image's caption. More precise region segmentation methods could possibly be used to better identify the marked areas. Another limitation of our current region placement strategy is that it does not consider the case of multiple markers denoting a single region. For example, authors commonly mark a single large region within an image by placing multiple arrows along the region's border. Instead of extracting a single visual ROI for such a region, our current method would produce a separate patch for each arrow. Such region placement considerations may significantly impact our classification method.

Our visual ROI classification approach performs reasonably well at labeling image patches with our five selected concepts. While it would have been preferable to evaluate our classifiers on a held-out data set with a greater number of concepts, we lacked a sufficiently large number of paired ROIs for both training and testing purposes. We hope to add additional concepts to our ontology in the future. The performance of our classifiers on the unfiltered set of paired ROIs produced by our actual textual and visual ROI extraction methods is especially encouraging considering they are sensitive to the combined errors of both methods. Moreover, our ROI descriptions do not always correspond neatly to one of our five chosen concepts. For example, the description "cystic bronchiectasis" corresponds to both our "cyst" and "bronchiole" classes. For classification purposes, we currently label ROI descriptions with all their relevant concepts. Despite these considerations, our current results are promising within our larger goal of mapping the appearance of regions within images to semantically meaningful concepts. The reduction of errors in our ROI extraction methods and the use of additional content-based features may aid in improving the performance of our classifiers.

CONCLUSION

Information processing systems commonly target the content of images contained in biomedical articles for information extraction and retrieval. Unfortunately, the meaning of images cannot be understood by analyzing their content alone. The text describing them must also be considered. In order to improve the integration of textual and visual information, we described in this paper our initial progress towards creating a visual ontology for biomedical imaging. A visual ontology defines a set of visual entities, the relationships among them, and maps their appearance to textual concepts. Thus, it creates a bridge between the visual characteristics of image regions and their semantic interpretation.

We focused our proof-of-concept experiments in populating a visual ontology on thoracic CT scans and their captions taken from a collection of biomedical articles. Our methods encompassed a variety of rule-based and statistical natural language and image processing techniques in order to label image regions with appropriate descriptions taken from their captions. Although the scope of our current work was limited, we expect our methods to be generalizable to other imaging modalities provided the images contain markers that can be paired with textual descriptions. Lacking such information, our methods could be adapted for the interactive extraction and annotation of regions of interest.

To demonstrate the utility of a visual ontology, we trained a classifier to automatically assign thoracic imaging concepts to image regions based solely on their appearance. We evaluated our methods on a gold standard set of annotated image regions and descriptions that we created. While our current system is experimental, our results demonstrated that our methods are encouraging first steps towards the creation of a visual ontology of biomedical imaging entities.

ACKNOWLEDGEMENTS

This work was supported by appointments to the NLM Research Participation Program and the intramural research program of the U. S. National Library of Medicine, National Institutes of Health.

REFERENCES

1. Simpson MS, Demner-Fushman D. Biomedical text mining: A survey of recent progress. In: Aggarwal CC, Zhai C, editors. Mining Text Data. Springer; 2012, p. 465–517.
2. Demner-Fushman D, Antani SK, Simpson MS, Rahman MM. Combining text and visual features for biomedical information retrieval. Tech. Rep.; Lister Hill National Center for Biomedical Communications; 2010.
3. Krishnam MS, Suh RD, Tomasian A, Goldin JG, Lai C, Brown K, et al. Postoperative complications of lung transplantation: Radiologic findings along a time continuum. Radiographics 2007;27(4):957–74.
4. Langlotz CP. RadLex: A new method for indexing online educational materials. Radiographics 2006;26(6):1595–7.
5. Yu H, Agarwal S, Johnston M, Cohen A. Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension. J Biomed Discov Collab 2009;4(1):1.
6. Hersh WR, Müller H, Jensen JR, Yang J, Gorman PN, Ruch P. Advancing biomedical image retrieval: Development and analysis of a test collection. J Am Med Inform Assoc 2006;13(5):488–96.
7. Müller H, Clough P, Deselaers T, Caputo B, editors. ImageCLEF: Experimental Evaluation in Visual Information Retrieval; vol. 32 of *The Information Retrieval Series*. Springer; 2010.
8. Lim JH, Chevallet JP. VisMed: A visual vocabulary approach for medical image indexing and retrieval. In: Lee G, Yamada A, Meng H, Myaeng S, editors. Information Retrieval Technology; vol. 3689 of *Lecture Notes in Computer Science*. Springer; 2005, p. 84–96.

9. Hollink L, Worring M. Building a visual ontology for video retrieval. In: Proceedings of the 13th annual ACM International Conference on Multimedia. 2005; p. 479–82.
10. Wei XY, Ngo CW. Ontology-enriched semantic space for video search. In: Proceedings of the 15th international conference on Multimedia. 2007; p. 981–90.
11. Jiang SQ, Du J, Huang QM, Huang TJ, Gao W. Visual ontology construction for digitized art image retrieval. J Comput Sci Tech 2005;20(6):855–60.
12. Park KW, Jeong JW, Lee DH. Olybia: Ontology-based automatic image annotation system using semantic inference rules. In: Kotagiri R, Krishna P, Mohania M, Nantajeewarawat E, editors. Advances in Databases: Concepts, Systems and Applications; vol. 4443 of *Lecture Notes in Computer Science*. Springer; 2007, p. 485–96.
13. Srihari RK, Burhans DT. Visual semantics: Extracting visual information from text accompanying pictures. In: Proceedings of the Twelfth National Conference on Artificial Intelligence. 1994; p. 793–8.
14. Mitchess M, Dodge J, Goyal A, Yamaguchi K, Stratos K, Han X, et al. Midge: Generating image descriptions from computer vision detections. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012; p. 747–756.
15. Müller H, Kalpathy-Cramer J, Eggel I, Bedrick S, Kahn CE Jr, Hersh W. Overview of the CLEF 2010 medical image retrieval track. In: Working Notes of CLEF 2010. 2010.
16. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: Glossary of terms for thoracic imaging. Radiology 2008;246(3):697–722.
17. Ide NC, Loane RF, Demner-Fushman D. Essie: A concept-based search engine for structured biomedical text. J Am Med Inform Assoc 2007;1(3):253–63.
18. Lindberg D, Humphreys B, McCray A. The unified medical language system. Methods Inf Med 1993;32(4):281–91.
19. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552–6.
20. Apostolova E, Demner-Fushman D. Towards automatic image region annotation: Image region textual coreference resolution. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. 2009; p. 41–4.
21. Klein D, Manning CD. Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003; p. 423–30.
22. You D, Antani S, Demner-Fushman D, Rahman MM, Govindaraju V, Thoma GR. Biomedical article retrieval using multimodal features and image annotations in region-based CBIR. In: Likforman-Sulem L, Agam G, editors. Document Recognition and Retrieval XVII; vol. 7534 of *Proceedings of SPIE*. 2010; p. 7534 0V.
23. You D, Antani S, Demner-Fushman D, Rahman MM, Govindaraju V, Thoma GR. Automatic identification of ROI in figure images toward improving hybrid (text and image) biomedical document retrieval. In: Agam G, Viard-Gaudin C, editors. Document Recognition and Retrieval XVIII; vol. 7874 of *Proceedings of SPIE*. 2011; p. 7874 0K.
24. Srinivasan GN, G S. Statistical texture analysis. In: Proceedings of World Academy of Science, Engineering and Technology; vol. 36. 2008; p. 1264–9.
25. Tamura H, Mori S, Yamawaki T. Textural features corresponding to visual perception. IEEE Trans Syst, Man, Cybern 1978;8(6):460–73.
26. Chatzichristofis SA, Boutalis YS. CEDD: Color and edge directivity descriptor – A compact descriptor for image indexing and retrieval. In: Gasteratos A, Vincze M, Tsotsos J, editors. Computer Vision Systems; vol. 5008 of *Lecture Notes in Computer Science*. Springer; 2008, p. 312–22.
27. Chatzichristofis SA, Boutalis YS. FCTH: Fuzzy color and texture histogram – A low level feature for accurate image retrieval. In: Ninth International Workshop on Image Analysis for Multimedia Interactive Services. 2008; p. 191–6.
28. Lux M, Chatzichristofis SA. LIRe: Lucene image retrieval – An extensible Java CBIR library. In: Proceedings of the 16th ACM International Conference on Multimedia. 2008; p. 1085–8.
29. Depeursinge A, Racoceanu D, Iavindrasana J, Cohen G, Platon A, Poletti PA, et al. Fusing visual and clinical information for lung tissue classification in high-resolution computed tomography. Artif Intell Med 2010;50(1):13–21.
30. Depeursinge A, Iavindrasana J, Hidki A, Cohen G, Geissbuhler A, Platon A, et al. Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization. J Digit Imaging 2010;23(1):18–30.
31. Olsson F, Eriksson G, Franzén K, Asker L, Lidén P. Notions of correctness when evaluating protein name taggers. In: Proceedings of the 19th International Conference on Computational Linguistics. 2002; p. 765–71.
32. Ooi GC, Khong PL, Chan-Yeung M, Ho JCM, Chan PKS, Lee JCK, et al. High-resolution CT quantification of bronchiectasis: Clinical and functional correlation. Radiology 2002;225(3):663–72.