The role of image modality and visual characteristics in archiving biomedical images

Sameer Antani, Daekeun You, Matthew Simpson, Mahmudur Rahman, Dina Demner-Fushman, George Thoma; National Library of Medicine, National Institutes of Health; Bethesda, MD

Abstract

Imaging in biomedicine has seen an explosive growth in recent decades. Clinicians can offer better diagnosis, and scientists and the lay public often better understand complex biomedical concepts through visual means. Typically, patient identifiers are used for archiving and indexing images' metadata in the clinical setting, and bibliographic citation data are used in library collections, such as the open access biomedical research articles from the U.S. National Library of Medicine's (NLM) PubMed Central® (PMC) repository. Automatically detected image modality and visual image characteristics can offer valuable addition to these traditional textual metadata for archiving and indexing visual material. Example image modalities include computerized tomography (CT), X-ray, magnetic resonance imaging (MRI), ultrasound, photographs, illustrations, charts, graphs, and sketches. The open-access biomedical literature in PMC is a source for OpenI (pronounced Open "eye"), a multimodal biomedical information retrieval system developed at the NLM that enables users to search for and retrieve relevant images and text. Our methods were evaluated in an international benchmarking forum in which we achieve a classification accuracy exceeding 90% at the highest level of a hierarchically organized image modality taxonomy, and 63.2% at the leaf level.

Introduction

The use of visual material in research publications has seen an explosive growth. Visual material is valuable not only in succinctly summarizing scientific and statistical results, but also expressing with high clarity what may take pages to describe. In particular in biomedicine, clinicians can offer better diagnosis, and scientists and the lay public are often able to better understand complex biomedical concepts through visual means. A study showed that figures are often among the first parts of articles that are reviewed by readers [1]. Therefore, it is of great significance if retrieval systems can provide the capability of searching for figures effectively.

We recently conducted an analysis of the available open access biomedical research articles from the U.S. National Library of Medicine's (NLM) PubMed Central[1]® (PMC) repository and found that currently 1.8 million images are available as a part of the freely downloadable data subset at the rate of about 3.5 figures per article. This is but a fraction of all images in the published biomedical literature. Analysis of the collection also reveals a sampling of the types (modalities) of images. These include those acquired through transmissive imaging techniques such as computerized tomography (CT), X-ray; nuclear imaging techniques, such as magnetic resonance imaging (MRI), positron emission tomography (PET); reflective imaging techniques, such as ultrasound and photography; and for scientific articles in particular, illustrative images, which includes charts, graphs, sketches, and so on.

In order to archive and index biomedical images, a typical method has been to use available metadata such as the patient identifier in the clinical setting, or, bibliographic citation information such as title, authors, abstract, keywords, journal title, publication data, etc., in a library or archive. Recently some indexing techniques have also started including text from figure captions as metadata. Examples include Yale Image Finder and The American Roentgen Ray Society's (ARRS) Goldminer[2]® system [2]. While these techniques are valuable they do not capture the visual characteristics of the image or its modality, which can be key distinguishing characteristics. Image modality and visual image characteristics can offer valuable addition to traditional metadata for archiving and indexing visual material (along with text material) in the biomedical literature. In previous work we showed that visual content of the images combined with the automatically computed image modality, figure caption text, and other traditional metadata can significantly improve indexing and retrieval of biomedical material over text retrieval alone [3].

Our methods for automatic multimodal image modality detection were evaluated in an international benchmarking forum called ImageCLEF[3] and achieved a high image modality classification accuracy and ranked within the submissions from the top three groups from among over 20 groups consisting of teams from academia and industry. In particular, we participated in ImageCLEFmed, the medical retrieval track of the benchmarking forum [4]. ImageCLEF is part of CLEF (the Cross Language Evaluation Forum) and has been organized yearly since 2003. It focuses on medical image annotation and retrieval and aims to provide support and resources for the evaluation of visual information retrieval systems. ImageCLEFmed 2012 consisted of three subtasks: modality classification, ad hoc image retrieval, and case-based article retrieval [5]. The modality classification task aims to evaluate the state of the art in figure classification.

The remainder of the article is organized as follows. We describe the modality hierarchy and classification strategy in

[1] http://www.ncbi.nlm.nih.gov/pmc/

[2] Goldminer: http://goldminer.arrs.org

[3] ImageCLEF: http://www.imageclef.org

Methods. In Evaluation we describe the performance of our techniques. In Discussion we underscore the value of using the image modality as metadata in archiving and indexing. Finally, we conclude and provide next steps in Conclusions.

Methods

This section describes the image modality hierarchy that is critical in organizing the classification strategy and computing the metadata for archiving the images. It also describes the data set and image features and has other relevant discussion on the topic.

Modality

The modality of an image can be computed for its appearance, and / or its function. In the broadest terms, image modality is the *type* of the image. However, when the images are organized, the taxonomy can take different forms. Through our continued participation in ImageCLEFmed, and our analysis of the open access biomedical literature in PMC, we use a modified hierarchical taxonomy based on one proposed for ImageCLEFmed 2012 [6]. The original taxonomy was modified after noting limitations in the benchmark hierarchy. The published taxonomy had a greater focus on image function that made using visual features for describing image content more challenging. Our modifications, instead, focus on image archiving, indexing, and retrieval. Our new hierarchy enables improved automatic image modality detection.

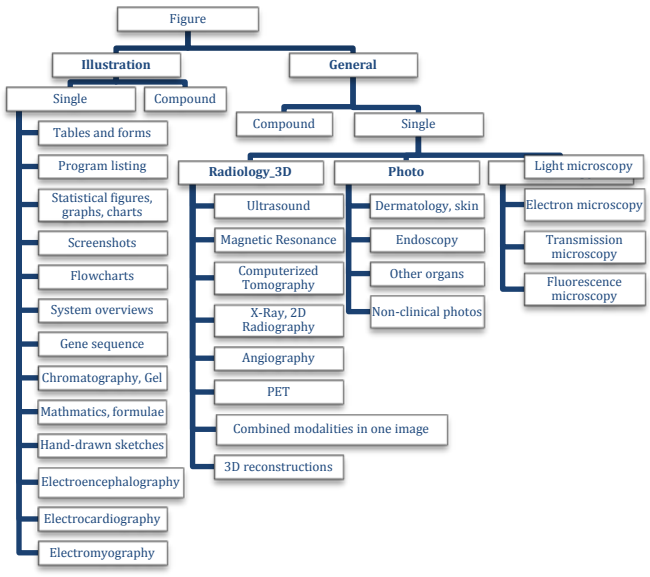

Figure 1. *Modified image modality taxonomy*

Our new modality taxonomy, shown in Figure 1, organizes images with a combination of appearance and function. At the highest level the images are separated into *Illustration* (appearance) and *General* (diagnostic function) images. Illustration images include statistical figures, such as charts and graphs, diagrams, schematics, and sketches. The sole diagnostic images that have been included in this set are signal and wave images, e.g. electrocardiograms, electroencephalographic images. General images include diagnostic ones that are further subdivided into radiological images, microscopy images, 3D reconstructions, and photos. This classification also recognizes that images appearing in the biomedical literature can be organized in single panel or multi-panel figures. Compound or multi-panel figures are detected within each category, and processed to separate the panels into individual images for meaningful classification [7]. These and other single panel images are then classified according to a mix of appearance, function, and acquisition.

Features

With the availability of the full text of the article including the figure captions, we took a multimodal classification approach that represents each image in the collection using text, visual, and combined features. Textual features are extracted from figure captions, discussion of the figure elsewhere in the article, and also citations listed as references. Visual features were computed from images and other figures. In our classification method, textual and visual features are extracted separately and then combined to achieve better classification performance.

Textual feature extraction

We represent each image as a structured document of image-related text, including the figure caption, discussion of the figure in the full text of the article, and other traditional bibliographic citation information. The structured documents may be indexed and searched with a traditional search engine, such as Lucene, or the underlying term vectors may be retrieved and added to a mixed image feature vector (multimodal, textual, and visual). To extract textual features for images, the terms in a structured document are commonly represented as a `bag of words'.

Visual feature extraction

From visual content we extract 15 features [8], including. color features such as the Color Layout Descriptor (CLD) of MPEG-7 [9]; edge features such as the Edge Histogram Descriptor (EHD); texture features such as moment-based features, wavelet, and Tamura feature; and features from combining two different types of features such as Color Edge Direction Descriptor (CEDD) and Fuzzy Color Texture Histogram (FCTH) available from the Lucene image retrieval engine (LIRE) library [10]. Each feature is extracted individually from the images and several features are combined into a single feature vector to represent an image.

Illustration class specific feature extraction

Illustration images in biomedical research articles significantly outnumber *General* images, as organized in Figure 1. Our analysis of the open access articles in PMC found that over 80% of the figures are illustrations. To increase the accuracy in classifying them, we implemented figure-type specific rule-based visual feature extraction methods. Illustrations frequently contain polygonal structures and overlaid graphic text strings on white background, e.g. flowcharts, program listings, schematics, tables, etc. We developed two key technologies for illustration figures: (i) a specialized optical character recognition (OCR) method for detecting text in illustrations; and (ii) a technique to recognize geometric patterns, such as dots, lines, and other polygonal shapes.

The former was developed in-house to recognize text in images, while the latter was developed using a combination of software modules from OpenCV[4], an open-source image processing and computer vision library. These techniques are used to extract polygonal shapes and text strings from images and used as features in the classification step. For example, computer program listings and algorithms mainly consist of text strings organized in a predictable pattern, while polygonal shapes are likely in figures with chemical structures.

In addition to visual features, we also extract keywords from the captions that can be used to recognize the figure type. For example, the caption may contain words or phrases such as "Table", "drawing of", "diagram of", "code", the chemical or gene name, and so on. Though these keywords and phrases appear to make unnecessary the extraction of visual features, there are many examples where these words could be misleading or absent, resulting in misclassification. The combination of visual and text features significantly improves figure-type classification accuracy.

Feature Selection

Not all attributes in a feature contribute to discrimination between image classes. Attribute (or feature) selection is the process of identifying those that do aid in the classification goal, thereby reducing the dimensionality of the feature vector, and the computational burden. We used the WEKA data-mining tool for this task. We also found that the Support Vector Machine (SVM) classifier performs best in our experiments.

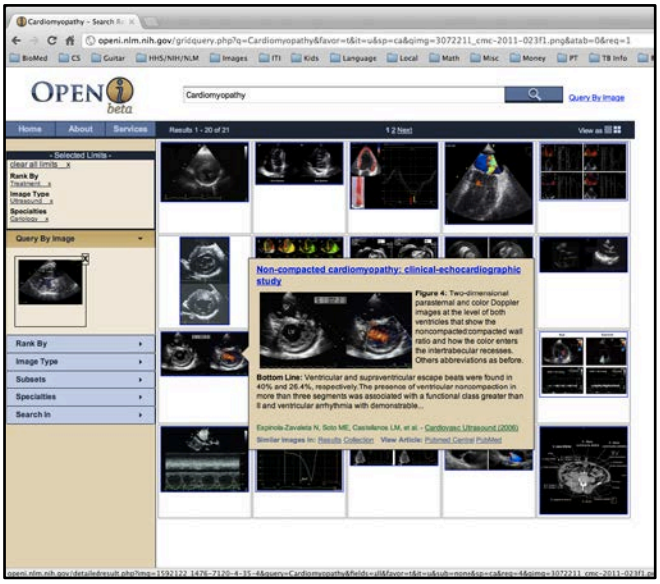

Figure 2. *OpenI grid view result screen showing query image on left, search term, and summary view of the enriched citation.*

[4] OpenCV: http://opencv.org

Evaluation

We evaluated the image modality classification on a data set of 2000 figures. The figures are sourced from the PMC and annotated with judgments from multiple observers and made available through the ImageCLEFmed forum. 1000 of these images were used for training the SVM classifier and 1000 were used in testing. In the evaluation we consistently found that combined features performed far better than either visual or text features. For example, the accuracy of recognizing illustration images is 96.28% versus 95.85% for visual features and 78.62% for text features. A similar separation occurs between radiology, microscopy, and photos where combined features are 93.53% accurate versus 87.42% and 83.76% for visual and text features, respectively.

We also determined that hierarchical image classification outperforms flat image modality organization in accuracy by 1.5%. In hierarchical classification each image is classified (and annotated) at each tier of the hierarchy. A separate classifier is trained at each level. For this experiment we developed six classifiers. Top classification performance was at 63.2% using combined features. Overall, visual features performed better than text features for image modality detection and combined features performed best overall.

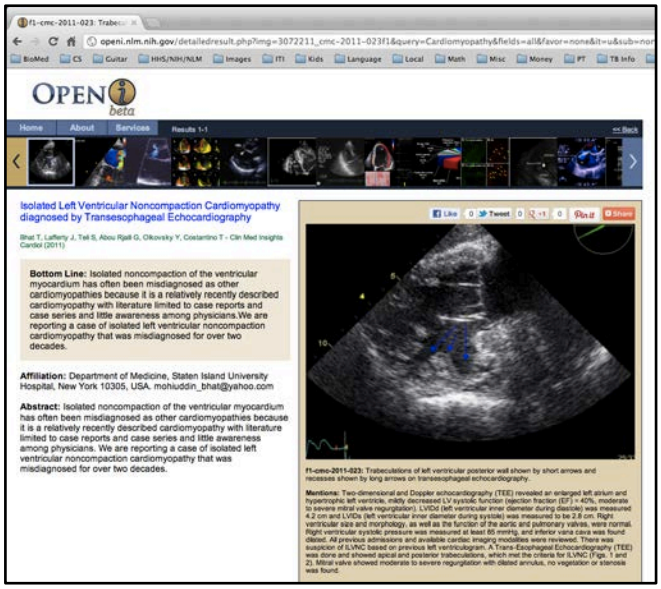

Figure 3. *OpenI enriched citation view showing result figure, its caption, bibliographic citation, and automatically extracted bottom line from the article.*

Discussion

Research literature is typically indexed and archived using traditional bibliographic information, viz., title, authors, publisher, publication details, keywords, and abstract. In addition for biomedical articles indexed in MEDLINE®, indexing terms include medical subject headings (MeSH®) terms that are keywords added by NLM. Traditional archives, indexing, and retrieval systems associate images with articles within which they occur. This is a logical representation and convenient for retrieval if the only mechanism available is using the text data listed above.

For repositories, such as PMC, and at publisher-hosted image repositories, such as SpringerImages[5], figure captions have also been recently added to the indexes allowing search for images using keywords associated with them.

The open access literature in PMC is the source for OpenI[6] (pronounced Open "eye"), a multimodal biomedical information retrieval system developed at NLM that enables users to search for and retrieve relevant images and text. In contrast to traditional approaches, OpenI indexes all the text and illustrations in medical articles in PMC by both textual and image-based features. The image features are used as inputs to a classifier to determine the image modality. The visual content in these images along with textual attributes, and the automatically computed image modality are used to compute a seamless multimodal index. The system, shown in Figure 2, can be queried with text words or by example images, and returns relevant images in a grid view or a list view. The user can click on an image to view, as shown in Figure 3, *enriched citations* that include the abstract and bibliographic citation (as from a traditional bibliographic system such as MEDLINE), but also enriched by relevant images and the figure caption. The user has the ability to filter images by 8 modalities and will soon be updated to the taxonomy shown in Figure 1 in upcoming releases. The system represents the next generation of indexing, archiving, and information retrieval, and question answering services with a goal to seamlessly retrieve relevant information from repositories without placing constraints on the modality of the query or retrieved objects. OpenI currently hosts 1.8 million images from 450,000 open access biomedical research articles from PMC. In less than a year after its public release, the site attracts over 15,000 unique visitors daily.

Conclusion

We have described the challenge in indexing and archiving images along with textual material. Metadata used for archiving biomedical research articles, for example, are content descriptors such as bibliographic citations and other keywords. In conventional archives or retrieval systems images have been included only as supplementary material. Such an organization does not support data reuse where deeper relationships between the visual data and the text can be derived and exploited.

We present our method for addressing this problem by developing a taxonomy of image appearance and use within the disciplines of clinical medicine and biomedical research. Our methods capture the visual content in the images and use that information along with any supporting text information to automatically determine the image modality. The modality information along with the visual content descriptors are used to index the images to aid multimodal information retrieval, data mining, clinical question answering, and clinical decision support applications.

[5] SpringerImages: http://www.springerimages.com
[6] OpenI: http://openi.nlm.nih.gov

Acknowledgements

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] R. J. Sandusky and C. Tenopir, Finding and using journal-article components: impacts of disaggregation on teaching and research practice, Journal of the American Society for Information Science and Technology, 59(6): 970-982, 2008.

[2] P. Ghosh, S. Antani, L. R. Long, G. R. Thoma, Review of medical image retrieval systems and future directions, 24th International Symposium on Computer-Based Medical Systems, pp. 1-6, Bristol, UK, June 2011.

[3] D. Demner-Fushman, S. Antani, M. Simpson, G. R. Thoma, Design and development of a multimodal biomedical information retrieval system, Journal of Computing Science and Engineering, 6(2):168-177, June 2012.

[4] M. S. Simpson, D. You, M. M. Rahman, D. Demner-Fushman, S. Antani, G. R. Thoma, ITI's participation in the ImageCLEF 2012 Medical Retrieval and Classification Tasks. *CLEF 2012 Working Notes, Rome, Italy, September 2012.*

[5] H. Müller, A. G. S. Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani, I. Eggel, Overview of the ImageCLEF 2012 Medical Image Retrieval and Classification Tasks. CLEF (Online Working Notes/Labs/Workshop) 2012.

[6] H. Müller, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani. Creating a Classification of Image Types In the Medical Literature For Visual Categorization. Proceedings of SPIE Medical Imaging: Advanced PACS-based Imaging Informatics and Therapeutic Applications. 8319:83190P-1-12, 2012.

[7] B. Cheng, S. Antani, R. J. Stanley, D. Demner-Fushman, G. R. Thoma. Automatic Segmentation of Subfigure Image Panels For Multimodal Biomedical Document Retrieval. Proceedings of SPIE Electronic Imaging Science and Technology, Document Retrieval and Recognition XVIII;7874:78740Z, 2011.

[8] M. Simpson, M. Rahman, S. Phadnis, D. Demner-fushman, S. Antani, G. Thoma, Text- and content-based approaches to image modality classi_cation and retrieval for the ImageCLEF 2011 medical retrieval track. CLEF (Notebook Papers/Labs/Workshop) (2011).

[9] S. F. Chang, T. Sikora, A. Purl. Overview of the MPEG-7 standard. IEEE Transactions on Circuits and Systems for Video Technology, 11, 688-695 (June 2001).

[10] M. Lux, S.A. Chatzichristo. Lire: lucene image retrieval: an extensible java CBIR library. Proceedings of the 16th ACM international conference on Multimedia , 1085-1088, ACM, New York, NY, USA (2008).

Author Biographies

Sameer Antani leads R&D in biomedical image informatics, image processing, multi-modal information retrieval, global health, mHealth, and next-generation publishing as a Staff Scientist at the U.S. National Library of Medicine. He received his M.Eng and PhD in computer science and engineering from the Pennsylvania State University. Dr. Antani is a senior member of SPIE, and a member of AMIA, IEEE, and IEEE Computer Society, and editorial board member of Elsevier Journal on Computers in Biology and Medicine.

Daekeun You received his MS in electronics engineering from the Sogang University, Korea (2003) and his PhD in computer science and engineering from University at Buffalo, SUNY (2011). Since then he has worked in the Communications Engineering Branch (CEB) at NLM/NIH as a postdoctoral fellow. His work has focused on the development of

algorithms for image analysis and processing for biomedical article indexing and retrieval.

Matthew S. Simpson received a BS (2004) degree in computer engineering from Clemson University and MS (2008) and PhD (2011) degrees in electrical and computer engineering from the University of Maryland, College Park. He is currently a postdoctoral fellow in the Communications Engineering Branch of the Lister Hill National Center for Biomedical Communications, a division of the US National Library of Medicine. His research interests include biomedical image retrieval and natural language processing.

Mahmudur Rahman received his PhD in Computer Science from Concordia University, Montreal, Canada (2008). He also received Post-Doctoral training (2008-2011) as a Research Fellow in the Communications Engineering Branch of the U.S. National Library of Medicine. Currently, he is working as a Research Engineer in the same branch of NLM at NIH. Dr. Rahman's research interests include multi-modal information retrieval, medical image annotation and retrieval, and statistical and interactive learning in image retrieval.

Demner-Fushman, MD, PhD, is a Staff Scientist at the Lister Hill National Center for Biomedical Communications, NLM, NIH, and leads research in information retrieval and natural language processing; providing clinical decision support through linking evidence to patients' data; and answering clinical and consumer health questions. Dr. Demner-Fushman is a Fellow of the American College of Medical Informatics, an editorial board member of JAMIA, and a founding member of the ACL Special Interest Group on biomedical natural language processing.

George Thoma is a Branch Chief at the U.S. National Library of Medicine, directing research and development in: the extraction of bibliographic data from medical articles, biomedical imaging, virtual books, a family reunification system for mass disasters, and multimedia-rich interactive publications. These projects appear at archive.nlm.nih.gov. He earned a B.S. from Swarthmore College, and the M.S. and Ph.D. from the University of Pennsylvania, all in Electrical Engineering. Dr. Thoma is a Fellow of the SPIE.