JCSE Journal of Computing Science and Engineering

Regular Paper
Journal of Computing Science and Engineering,
Vol. 6, No. 2, June 2012, pp. 168-177

# Design and Development of a Multimodal Biomedical Information Retrieval System

**Dina Demner-Fushman***, **Sameer Antani, Matthew Simpson, and George R. Thoma**
National Library of Medicine, Bethesda, MD, USA
**ddemner@mail.nih.gov, santani@mail.nih.gov, simpsonmatt@mail.nih.gov, gthoma@mail.nih.gov**

## Abstract
The search for relevant and actionable information is a key to achieving clinical and research goals in biomedicine. Biomedical information exists in different forms: as text and illustrations in journal articles and other documents, in images stored in databases, and as patients' cases in electronic health records. This paper presents ways to move beyond conventional text-based searching of these resources, by combining text and visual features in search queries and document representation. A combination of techniques and tools from the fields of natural language processing, information retrieval, and content-based image retrieval allows the development of building blocks for advanced information services. Such services enable searching by textual as well as visual queries, and retrieving documents enriched by relevant images, charts, and other illustrations from the journal literature, patient records and image databases.

**Category:** Convergence computing

**Keywords:** Multimodal biomedical information retrieval; Natural language processing; Content-based information retrieval; Image processing; Advanced information services

## I. INTRODUCTION

The importance of illustrations in scientific publications is well-established. In a survey of information needs of researchers and educators, Sandusky and Tenopir [1] found that scientific journal-article components, such as tables and figures, are often among the first parts of an article scanned or read by researchers. In addition, the survey participants indicated that having access to the illustrations (we use "images," "illustrations," and "figures" interchangeably when referring to visual material in our set of medical articles) prior to obtaining the whole publication would greatly enhance their search experience. In the biomedical domain, Divoli et al. [2] found that bioscience literature search systems, such as PubMed, should show figures from articles alongside search results, and that captions should be searched, along with the article title, metadata, and abstract. Simpson et al. [3] showed that, for the system presented in this paper, retrieval of case descriptions similar to a patient's case was significantly improved with the use of image-related text.

The clear need for a multimodal retrieval system on the one hand, and the sufficient maturity of the information retrieval (IR) and content-based image retrieval (CBIR) techniques on the other, motivated us to implement a prototype multimodal system (called OpenI) for advanced information services. These services should enable:

- Searching by textual, visual and hybrid queries
- Retrieving illustrations (medical images, charts, graphs, diagrams, and other figures)

Open Access http://dx.doi.org/10.5626/JCSE.2012.6.2.168 http://jcse.kiise.org

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Received** February 18 2012, **Revised** May 26 2012, **Accepted** May 27 2012
*Corresponding Author

Copyright © 2012. The Korean Institute of Information Scientists and Engineers pISSN: 1976-4677 eISSN: 2093-8020

**Fig. 1.** Textual search results in the list view of the OpenI system. The navigation panel on the left allows filtering and sorting search results along several facets. The pop-up window that appears on scrolling over the image provides a brief, but key, summary of the information in the article.

- Retrieving bibliographic citations, enriched by relevant images
- Retrieving from collections of journal literature, patient records, and independent image databases
- Linking patient records to literature and image databases, to support visual diagnosis and clinical decision making

This paper first presents an overview of the processes involved in preparing multimodal scientific articles for indexing and retrieval. It then briefly introduces a distributed architecture that allows for both processing the original documents in reasonable time (for example, extracting image features from thousands of images in just a few hours), and for real-time retrieval of the processed documents. The paper concludes with a discussion of the implemented system prototype (Fig. 1) that currently provides access to over 600,000 figures from over 250,000 medical articles, evaluation of the algorithms constituting the system, and future directions in multimodal retrieval and its evaluation.

## II. BUILDING BLOCKS FOR ADVANCED INFORMATION SERVICES

To prepare documents for indexing and retrieval, we combine our tools and those publicly available, in a pipeline that starts with acquiring data and ends in the generation of MEDLINE citations enriched with image-related information (henceforth, "enriched citations"). The initially separate text and image processing pathways merge to create multimodal indexes, for use with specialized multimodal information retrieval algorithms in Fig. 2. The images and text data in the current prototype are obtained from the open access subset of PubMed Central (PMC), using the PMC file transfer protocol (FTP) services (http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/). This

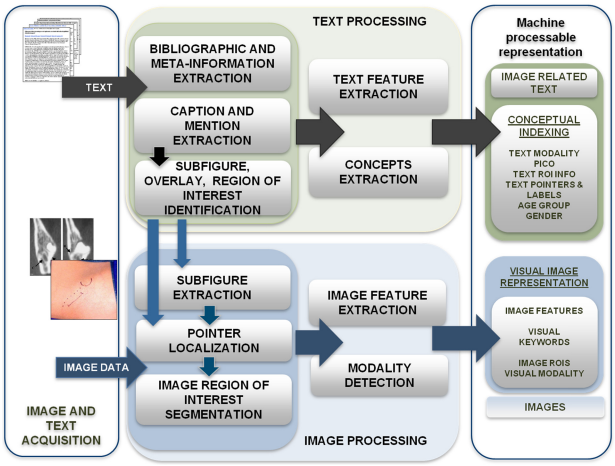


**Fig. 2.** Overview of image and text processing steps for creating enriched citations. The text processing module (see Section III) extracts descriptions of images and image captions from the full text articles, to enrich the MEDLINE citation of the article containing the image. The image processing module (see Section IV) extracts the low-level visual features used in image modality classification and image clustering. The image clusters are labeled with alphanumeric strings ("cluster words"). Subsequently, image features are represented using the cluster words. The cluster words pertaining to an image are added to its enriched citation, along with the image modality label. An example enriched citation is shown in Fig. 3.

full-text archive of biomedical publications provides the text of each article in extensible markup language (XML) format, and all published figures as JPEG files. The XML files serve as input to the modules assembled in the text processing pipeline, and the images are processed through the image processing pipeline. Several image processing modules (for example, modality classifiers) require output of the text processing modules as additional input.

The output of the document processing pipeline is a set of enriched MEDLINE citations in XML format that is subsequently indexed with the National Library of Medicine (NLM)'s domain-specific search engine Essie [4], as well as with the widely-used open-source search engine Lucene (http://lucene.apache.org/). An enriched citation consists of three parts: 1) the original bibliographic citation obtained using E-Utilities (http://www.ncbi.nlm.nih.gov/corehtml/query/static/eutils_help.html); 2) the image caption and image-related paragraphs extracted from the full-text of the article, along with salient information extracted from this free text and stored in structured form; and 3) image features expressed as searchable alphanumeric strings, along with the image uniform resource identifier (URI), for display in the user interface.

Some of the processing steps (such as extracting elements of an XML document) are well-known, and will be omitted here. We will focus instead on the overall process flow, and the unique challenges and opportunities presented by images found in biomedical publications.

### A. Generating an Enriched Citation

The OpenI document processing system is developed in Java, and uses Hadoop MapReduce (Apache Software Foundation, Los Angeles, CA, USA) to parallelize text processing and image feature extraction. An enriched citation object is generated in the text processing pipeline, which is presented in Section III. Images are processed independently, and the information extracted from the images is added to the enriched citation in the final merging step.

One challenge in image processing arises from several illustrations combined into one figure. These multi-panel (or compound) images found in many articles reduce the quality of image features, if the features are extracted from the whole image. For feature extraction, therefore, these images need to be first separated into distinct panels. This process is described in Section V.

In addition to the text and image features necessary for retrieval, each enriched citation also contains meta-information derived from the basic features (such as the medical terms found in the captions and mapped to the unified medical language system [UMLS] [5] concepts). This meta-information is used to filter and re-rank search results. For example, the results could be restricted to radiology images using the modality classification results, or re-ranked to promote articles focused on genetics (identified as such by genetics-related concepts in the titles, Medical Subject Heading [MeSH] terms, and captions). The currently available filters are described in Section VI.

## III. TEXT PROCESSING

The text processing begins with the extraction of the image caption and the paragraph(s) discussing the figure ("mentions"). In the PMC documents, captions are a defined XML element. We extract the mentions using regular expressions: we first find an indicator that a figure is mentioned, usually, words "Figure" or "Fig" (sometimes within mark-up tags or punctuation) followed by a number, and then extract the paragraph around the indicator.

Next, the caption processing module determines if the caption belongs to a multi-panel figure. The rule-based system is looking for sequences of alphanumeric characters that are included within repeating tags, or followed by a repeating punctuation sign (for example, A. B. C.) If a sequence is found, the number of panels and the panel labels are added to the enriched citation.

The next module extracts the descriptions of image overlays (such as arrows), and regions of interest (ROI) indicated by the overlays [6]. The ROI descriptions added to the enriched citations are currently a searchable field. The whole output of the module is needed for our ongoing research, in building a visual ontology that will associate the UMLS concepts with specific image features.

Finally, a concept extraction module submits the captions and mentions to MetaMap [7], which identifies UMLS concepts in the text. The module then applies rules and stop-word lists to the MetaMap output, to reduce the set of the identified UMLS concepts to the salient disorder, intervention and anatomy terms [8]. Fig. 3 shows an enriched citation in XML format.

## IV. IMAGE PROCESSING

Low-level visual features, such as color, texture, and shape, are insufficient for capturing image semantics, but they are the primary building blocks of the visual content in an image. They can be effective, if a judiciously selected feature metric is used to capture the visual content in an image, and then incorporates it into a suitable machine-learning framework that supports multi-scalar and concept-sensitive visual similarity. In the OpenI prototype system, the low-level visual features extracted from the whole set of images are first clustered using the k-means algorithm, and then the resulting clusters are labeled with alphanumeric strings ("cluster words") that serve as cluster identifiers. The images are then annotated with the cluster identifier of each low-level feature. These cluster words are added to the enriched citation in Fig. 3, as the last document preparation step. The enriched citation field

```xml
  <?xml version="1.0" encoding="utf-8" ?>
- <document>
    <meta iclef_id="239029" />
    <meta publisher="Radiology" />
    <meta journal_title="The puff of smoke sign" />
    <meta fulltext_html_url="http://radiology.rsnajnls.org/cgi/content/full/247/3/910" />
    <meta iti_id="18487544F1" />
    <meta volume="247" />
    <meta authors="Ortiz-Neira, Clara L;" />
    <meta pmid="18487544" />
    <meta issue="3" />
    <title>The puff of smoke sign</title>
    <abstract />
  - <image type="figure" id="1" src="./images/239029.jpg"
      link="http://radiology.rsnajnls.org/cgi/content/full/247/3/910/F1">
      <caption>Anteroposterior angiogram of right internal carotid artery shows abnormal hypertrophy of
        perforating arteries, which produces the puff of smoke sign (arrow) and is associated with
        narrowing (arrowheads) of the M1 and A1 segments of the distal internal carotid artery.</caption>
      <mention />
    - <pico>
        <modalityclass>xr</modalityclass>
        <modality>angiogram</modality>
        <intervention cui="C0002978" negstatus="NOT_NEGATED">angiogram</intervention>
        <anatomy cui="C0226156" negstatus="NOT_NEGATED">right internal carotid artery</anatomy>
        <problem cui="C0020564" negstatus="NOT_NEGATED">hypertrophies</problem>
        <anatomy cui="C1182750" negstatus="NOT_NEGATED">perforating arteries</anatomy>
        <anatomy cui="C0007276" negstatus="NOT_NEGATED">internal carotid artery</anatomy>
      </pico>
    - <rois>
        <roi type="arrow">smoke sign</roi>
        <roi type="arrow">narrowing</roi>
      </rois>
    </image>
  + <mesh>
  </document>
```


**Fig. 3.** Enriched MEDLINE citation.

containing cluster words is indexed in the same manner as the rest of the citation.

## A. Feature Extraction

Images in the open access PMC collection are of different sizes. In order to obtain a uniform measure with greater computational efficiency, we compute features from images that are reduced to a common size, measuring 256 × 256 pixels. In the future, we intend to process images at a significantly higher (or full) resolution, to extract meaningful local features.

### *1) Color Features*

Color plays an important role in the human visual system, and measuring its distribution can provide valuable discriminating data about the image. We use several color descriptors to represent the color in the image. To represent the spatial structure of images, we utilize the color layout descriptor (CLD) [9] specified by MPEG-7 [10]. The CLD represents the spatial layout of the images in a compact form, and can be computed by applying the discrete cosine transformation (DCT) to the 2D array of local representative colors in the YCbCr color space, where Y is the luminance component, and Cb and Cr are the blue and red chrominance components, respectively. Each color channel is 8-bits, and is represented by an average value computed over 8 × 8 image blocks. We extract a CLD with 10 Y, 3 Cb, and 3 Cr components, to form a 16-dimensional feature vector.

Another feature used is the color coherence vector (CCV) [11] that captures the degree to which pixels of that color are members of large similarly colored regions. A CCV stores the number of coherent versus incoherent pixels with each color, thereby providing finer distinctions than color histograms. Color moments, also computed in the perceptually linear L*a*b* color space, are measured, using the three central color moment features: mean, standard deviation, and skewness. Finally, 4 dominant colors in the standard red, green, blue (RGB) color space and their degrees are computed, using the k-means clustering algorithm.

### *2) Edge Features*

Edges are not only useful in determining object outlines, but their overall layout can be useful in discriminating between images. The edge histogram descriptor (EHD) [9], also specified by MPEG-7, represents a spatial distribution of edges in an image. It computes local edge distributions in an image, by dividing the image into 4 × 4 sub-images, and generating a coarse-orientation histogram from the edges present in each of these sub-images. Edges in the image are categorized into five types: vertical, horizontal, 45 × diagonal, 135 × diagonal, and other

non-directional edges. A finer-grained histogram of edge directions (72 bins of 5 × each) is also constructed from the output of a Canny edge detection algorithm [12] operating on the image. This feature is made invariant to image scale, by normalizing it with respect to the number of edge points in the image.

***3) Texture Features***

Texture measures the degree of "smoothness" (or "roughness") in an image. We extract texture features from the four directional gray-level co-occurrence matrices (GLCM) that are computed over an image. Normalized GLCMs are used to compute higher order features, such as energy, entropy, contrast, homogeneity and maximum probability. We also compute Gabor filters to capture image gist (coarse texture and spatial layout). The gist computation is resistant to image degradation, and has been shown to be very effective for natural scene images [13]. Finally, we use the discrete wavelet transform (DWT) that has been shown to be useful in multi-resolution image analysis. It captures image spatial frequency components at varying scales. We compute the mean and standard deviation of the magnitude of the vertical, horizontal, and diagonal frequencies at three scales.

***4) Average Gray Level Feature***

This feature is extracted from the low-resolution scaled images, where each image is converted to an 8-bit gray-level image, and scaled down to 64 × 64 pixels, regardless of the original aspect ratio. Next, this reduced image is partitioned further with a 16 × 16 grid, to form small blocks of (4 × 4) pixels. The average gray value of each block is measured and concatenated, to form a 256-dimensional feature vector.

***5) Other Features***

We extract two additional features using the Lucene image retrieval engine (LIRE) library: the color edge direction descriptor (CEDD) and the fuzzy color texture histogram (FCTH) [14]. CEDD incorporates color and texture information into a single histogram, and requires low computational power compared to MPEG-7 descriptors. To extract texture information, CEDD uses a fuzzy version of the five directional edge filters used in MPEG-7 EHD that were described previously. This descriptor is robust with respect to image deformation, noise, and smoothing. The FCTH uses fuzzy high frequency bands of the Haar wavelet transform to extract image texture.

## V. MULTI-PANEL FIGURE SEGMENTATION

Independent of the caption processing module that outputs the number of panels and panel labels, two image-feature based modules detect panel boundaries and panel labels. The output of the three modules is used in the panel splitting module.

The image-feature based panel segmentation module determines if an image contains homogeneous regions that cross the entire image. If no homogeneous regions are found, the image is classified as single-panel. If homogeneous regions are found, the panel segmentation module iteratively determines if each panel contains homogenous regions, and finally outputs coordinates of each panel.

The label detection module [15] first binarizes the image into black and white pixels, then, searches the image for connected components that could represent panel labels, and then applies optical character recognition (OCR) methods to the connected components. Finally, the most probable label sequences and locations are selected from all candidate labels, using Markov random field modeling.

The label splitting module takes the outputs of the caption splitting, panel segmentation and label detection modules, and splits the original figure if the following conditions are met: all three modules agree on the number of panels, and the caption splitting and label detecting modules agree on the labels (this happens for approximately 30% of the multi-panel figures). If the panel segmentation or the label detection modules fail completely, the image cannot be split. However, if the modules partially agree on labels, and position some of the labels at the corners of the corresponding panels, heuristics help to compensate for the partial errors of individual modules, and the combined information helps correctly split another 40% of the multi-panel images. The images on which the algorithm fails are processed as single-panel images. All images are displayed in the original form.

## VI. OPENI SYSTEM ARCHITECTURE

OpenI uses VMware vSphere 4 (VMware Inc., Palo Alto, CA, USA) and a high-performance Linux based storage area network (SAN) to support a fault-tolerant, scalable, and efficient production-grade system. A high-performance SAN storage cluster that was built using Red Hat Enterprise Linux (Red Hat, Raleigh, NC, USA) provides OpenI with a dedicated, reliable, predictable and high-performing storage system for the Hadoop cluster. VMware vSphere 4 is used to virtualize the Hadoop Namenode, and run it in fault-tolerant mode. The fault-tolerance feature of VMware allows a single virtual machine to run simultaneously on two hardware servers. Further, vSphere monitors the heartbeat of the Namenode, and restarts it automatically if it should become inoperable. These features eliminate the single point of failure, and turn Hadoop into a stable and reliable development and research platform. Finally, vSphere is used to make OpenI processes run efficiently on multi-core CPUs. Each compute node running processes that are not multi-core aware can be created as a virtual machine tied to a specific physical core.

## VII. OPENI IMAGE RETRIEVAL SYSTEM

The OpenI prototype (http://openi.nlm.nih.gov/) currently supports image retrieval for textual, visual and hybrid queries. The images submitted as queries are processed as described above, and represented using cluster "words". After this processing step, the cluster "words" are treated as any other search terms.

Based on the principles developed by [16], the search results are displayed either on a grid that allows a view of all top 20 retrieved images (Fig. 4), or as a traditional list in Fig. 1. In either layout, scrolling over the image brings up a pop-up window that, along with the traditional elements of search results, such as titles and author names, provides captions of the retrieved images, and short summaries of the articles.

**Fig. 4.** Search results in a grid display.

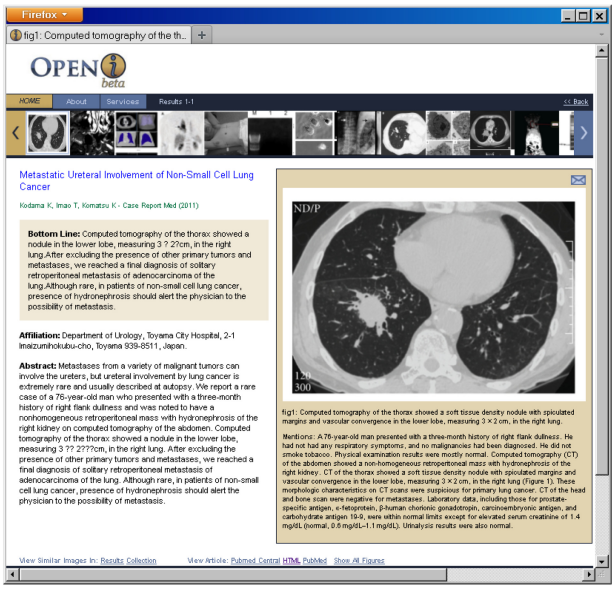


**Fig. 5.** A view of an enriched citation in the user interface. The ribbon at the top allows rapid navigation to other images in the search results. The links at the bottom allow to link out to the publisher's site, PubMed or PubMed Central, and search for similar images. Enriched citations can be sent using the email icon.

The summaries are obtained through RIDeM services (http://clinicalreferences.nlm.nih.gov/ridem). The summaries (or "bottom-line" patient-oriented outcomes extracted from abstracts) are generated by extracting three sentences most likely to discuss patient-oriented outcomes of the methods presented in the paper [17]. The probability of a sentence to be an outcome is determined by a meta-classifier that combines outputs of several base classifiers (such as a Naïve Bayes classifier, and classifiers based on the position of the sentence in the abstract, or the presence of relevant terms in the sentence).

Once the search results are displayed, the users can find similar images in the results and in the entire collection, by following links that perform new search requests for images visually similar to the currently selected image. These new searches are based purely on image features. Users may view all other images in a given paper without leaving the initial search page, and drill down to the full enriched citation in Fig. 5.

The search results can also be filtered using the following facets: 1) image type; 2) subsets of publications, for example, systematic reviews; 3) clinical specialties; 4) enriched citation fields.

### *A. Image Type*

The image type filter is based on our classification of images into eight medical images modalities, such as magnetic resonance imaging (MRI), x-ray, computed tomography (CT), and ultrasound. Our method [18] uses a support vector machine (SVM) to classify images into multiple modality categories. The degree of membership in each category can then be used to compute the image modality. In its basic formulation, the SVM is a binary classification method that constructs a decision surface, and maximizes the inter-class boundary between the samples. To extend it to multi-class classification, we combine all pair-wise comparisons of binary SVM classifiers, known as one-against-one or pair-wise coupling (PWC). Each SVM is trained for one image feature. The class with the greatest estimated probability for each feature accumulates one vote. The class with the greatest number of votes after classifying for all features is deemed to be the winning class, and the modality category of the class is assigned to the image. When a user requests a specific image type, a hard constraint on exactly matching the image modality field of the enriched citation is imposed.

### *B. Subsets*

Due to the nature of the collection, not all MEDLINE/PubMed subsets (http://www.nlm.nih.gov/bsd/pubmed_

subsets. html), such as the core clinical journals subset (http://www.nlm.nih.gov/bsd/aim.html), are available in OpenI. We used the subject field of the NLM's List of Serials Indexed for Online Users (http://wwwcf.nlm.nih.gov/serials/journals/) to categorize the journals into clinical specialties and subsets. Not all journals are assigned to broad subject areas, and only publications in journals that are assigned to an area will be retrieved, when a user requests to filter the results by a subset or specialty. Where available, we used the subset field of the original MEDLINE citation.

### C. Enriched Citation Fields

The users can search the text in any combinations of the following: titles, abstracts, captions, mentions, MeSH terms, and author names.

Finally, the search results could also be re-ranked according to the users' interests (indicated by selecting advanced search options) along the following axes: 1) by the date of publication (most recent or oldest first); 2) by the clinical task that is discussed in the paper (diagnosis, cause of the problem, prevention, prognosis, treatment, etc.). The clinical task that was the focus of the study presented in a paper is determined using rules that take into account specific MeSH terms. For example, the term "Infectious Disease Transmission, Vertical/*Prevention & Control" indicates that the task is prevention. The star indicates that the publication is focused on prevention.

## VIII. RELATED WORK

Several ongoing research efforts are dedicated to augmenting text results with images. Some of these efforts aim to retrieve images by matching query text terms in the citations to the articles and the figure captions. We list five efforts related to our goals. Most systems do not use image features to find similar images or combine visual and text features for biomedical information retrieval. Our goals include improving the relevance of multi-modal (text and image) information retrieval, by including lessons learned from these efforts.

The BioText [16] search engine searches over 300 open access journals, and retrieves figures, as well as text. BioText uses Lucene to search full-text or abstracts of journal articles, as well as image and table captions. Retrieved results (displayed in a list or grid view) can be sorted by date or relevance. This search engine has influenced our user interface design.

Yottalook (http://www.yottalook.com/) allows multilingual searching to retrieve information (text or medical images) from the Web and journal articles. The goal of the search engine is to provide information to clinicians at the point of care. The results can be viewed as thumbnails or details. This site sets an example in the breadth of its searches, capabilities to filter results on image modality and other criteria, being current with social media, and connecting with the users' myRSNA accounts (offered by the Radiological Society of North America [RSNA]), which allow saving of search results.

Other related work includes the Goldminer (http://goldminer.arrs.org/home.php) search engine developed by the American Roentgen Ray Society (ARRS) that retrieves images by searching figure captions in the peer-reviewed journal articles appearing in the RSNA journals, Radiographics and Radiology. It maps keywords in figure captions to concepts from the UMLS. Users have the options to search by age/modality/sex for images, where such information is available. Results are displayed in a list or grid view.

The FigureSearch (http://figuresearch.askhermes.org/articlesearch/index.php?mode=figure) system uses a supervised machine-learning algorithm for classifying clinical questions, and Lucene for information retrieval over the published medical literature, to generate a list view of the results with relevant images, abstracts, and summaries.

The Yale image finder (YIF) [19] searches text within biomedical images, captions, abstracts, and titles, to retrieve images from biomedical journal papers. YIF uses optical character recognition to recognize text in images, in both landscape and portrait modes.

The image retrieval in medical applications (IRMA) system (http://www.irma-project.org) primarily uses visual features and a limited number of text labels that describe the anatomy, biosystem, imaging direction, and modality of the image for medical image retrieval. We have collaborated with the developers of the IRMA system, and enhanced their image retrieval system (which uses features computed on the gross image), with our image features and similarity computation techniques applied to local image regions [20].

Increasing commercial interest in multi-modal information retrieval in the biomedical domain is indicated by the industry teams participating in the ImageCLEFmed (http://www.imageclef.org/2012/medical) evaluations dedicated to retrieval of medical images and similar patients' cases. Participants include researchers from Siemens, GE Medical Systems, Xerox, and other industrial organizations. Publishers such as Springer also provide text-based image retrieval (http://www.springerimages.com/). Other commercial image search engines include those developed by Google, Gazopa, and Flickr.

## IX. EVALUATION

Saracevic [21] defines six levels of evaluation of information retrieval (IR) systems: 1) the engineering level addresses speed, integrity, flexibility, computational effectiveness, etc.; 2) the input level evaluates the document collection, indexed by the system and its coverage; 3) the

processing level assesses performance of algorithms; 4) the output level evaluates retrieval results and interactions with the system; 5) the use and user level evaluates system performance for a given task; and 6) the social level evaluates the impact of the system on research, decision-making, etc.

OpenI is a complex system that has been evaluated along several of the aforementioned axes: on engineering, processing, and output levels. Its modality classification, ad-hoc image retrieval and case-based retrieval have been evaluated within the ImageCLEFmed evaluations since 2007. The OpenI results are steadily in the leading group, achieving, for example, 92% accuracy in the modality classification task in 2010, and 0.34 mean average precision in case-based retrieval in 2009.

In addition to testing the overall retrieval performance of the system, we have evaluated the benefits of enriching MEDLINE citations with image captions and passages pertaining to images [3]. As mentioned above, these passages significantly improved case-based retrieval. Other evaluations of the parts of the document preparation steps include evaluation of multi-panel image segmentation, and evaluation of the processes involved in automatic generation of a visual ontology, such as the ROI and ROI marker extraction from text and images. Overall, the panel splitting module achieves 80.92% precision at 73.39% recall. The ROI marker extraction from text achieves 93.64% precision and 87.69% recall, whereas ROI identification is at 61.15% accuracy [6]. The ROI marker extraction (limited to arrows) ranges from 75% to 87% accuracy for different arrow types [22].

User level evaluations are often approximated with site visits and click-through data [23]. To that end, we are monitoring the numbers of unique visitors per day, which for May 2012 is at approximately 700 on average, and shows a growing trend. We are also planning a social level experiment (with members of particular research and clinical communities) to evaluate the effectiveness of OpenI in assisting with specific tasks.

## X. FUTURE WORK

The deployment of the prototype system and the architecture presented in this paper allows continuing research in several directions. First, we are interested in the usability of the current user interface, and the usefulness of the search features. Second, we are expanding and improving the extraction of the basic image features, and the selection of these features for various higher-level tasks, such as image modality classification, image ROI recognition, and building a visual ontology. The latter task includes associating specific image features with the UMLS concepts. Recognizing that many researchers would like to focus on specific aspects of image retrieval, for example improving retrieval methods or text understanding, we plan to provide our current document processing methods as publicly available services.

## ACKNOWLEDGMENTS

This work would have not been possible without our amazing team of post-doctoral researchers and the equally talented OpenI staff. Our thanks go to Emilia Apostolova, Michael Chung, Glenn Ford, Michael Kushnir, Srinivas Phadnis, Md. Rahman, Daekeun You, and Zhiyun Xue. We thank Essie developers Russell Loane and Nicholas Ide for providing the search engine, and being wonderfully responsive and supportive in development of the OpenI systems.

This work was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

## 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*, vol. 59, no. 6, pp. 970-982, 2008.
2. A. Divoli, M. A. Wooldridge, and M. A. Hearst, "Full text and figure display improves bioscience literature search," *PLoS One*, vol. 5, no. 4, p. e9619, 2010.
3. M. S. Simpson, D. Demner-Fushman, and G. R. Thoma, "Evaluating the importance of image-related text for ad-hoc and case-based biomedical article retrieval," *American Medical Informatics Association (AMIA) Annual Symposium Proceedings*, vol. 13, pp. 752-756, 2010.
4. N. C. Ide, R. F. Loane, and D. Demner-Fushman, "Essie: a concept-based search engine for structured biomedical text," *Journal of the American Medical Informatics Association*, vol. 14, no. 3, pp. 253-263, 2007.
5. D. A. Lindberg, B. L. Humphreys, and A. T. McCray, "The unified medical language system," *Methods of Information in Medicine*, vol. 32, no. 4, pp. 281-291, 1993.
6. E. Apostolova and D. Demner-Fushman, "Towards automatic image region annotation: image region textual coreference resolution," *Proceedings of Human Language Technologies: the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, CO, 2009, pp. 41-44.
7. A. R. Aronson and F. M. Lang, "An overview of MetaMap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229-236, 2010.
8. D. Demner-Fushman, J. G. Mork, S. E. Shooshan, and A. R. Aronson, "UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text," *Journal of Biomedical Informatics*, vol. 43, no. 4, pp. 587-594, 2010.
9. S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: color and

edge directivity descriptor: a compact descriptor for image indexing and retrieval," *Computer Vision Systems*, *Lecture Notes in Computer Science vol. 5008*, A. Gasteratos, M. Vincze, J. K. Tsotsos, eds., Heidelberg: Springer Berlin, pp. 312-322, 2008.
10. S. F. Chang, T. Sikora, and A. Purl, "Overview of the MPEG-7 standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 688-695, 2001.
11. G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," *Proceedings of the 4th ACM International Conference on Multimedia*, Boston, MA, 1996, pp. 65-73.
12. J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679-698, 1986.
13. A. Oliva and A. Torralba, "Building the gist of a scene: the role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23-36, 2006.
14. M. Lux and S. A. Chatzichristofis, "LIRe: lucene image retrieval: an extensible java CBIR library," *Proceedings of the 16th ACM International Conference on Multimedia*, Vancouver, BC, Canada, 2008, pp. 1085-1088.
15. D. You, S. Antani, D. Demner-Fushman, V. Govindaraju, and G. R. Thoma, "Detecting figure-panel labels in medical journal articles using MRF," *Proceedings of 2011 International Conference on Document Analysis and Recognition*, Beijing, China, 2011, pp. 967-971.
16. M. A. Hearst, A. Divoli, and H. Guturu, A. Ksikes, P. Nakov, M. A. Wooldridge, and J. Ye, "BioText search engine: beyond abstract search," *Bioinformatics*, vol. 23, no. 16, pp. 2196-2197, 2007.
17. D. Demner-Fushman, B. Few, S. E. Hauser, and G. R. Thoma, "Automatically identifying health outcome information in MEDLINE records," *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 52-60, 2006.
18. M. M. Rahman, S. K. Antani, R. L. Long, D. Demner-Fushman, and G. R. Thoma, "Multi-modal query expansion based on local analysis for medical image retrieval," *Proceedings of the 1st MICCAI International Conference on Medical Content-Based Retrieval for Clinical Decision Support*, London, UK, 2009, pp. 110-119.
19. S. Xu, J. McCusker, and M. Krauthammer, "Yale image finder (YIF): a new search engine for retrieving biomedical images," *Bioinformatics*, vol. 24, no. 17, pp. 1968-1970, 2008.
20. S. K. Antani, T. M. Deserno, L. R. Long, and G. R. Thoma, "Geographically distributed complementary content-based image retrieval systems for biomedical image informatics," *Studies in Health Technology and Informatics*, vol. 129, pp. 493-497, 2007.
21. T. Saracevic, "Evaluation of evaluation in information retrieval," *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, 1995, pp. 138-146.
22. D. You, S. Antani, D. Demner-Fushman, M. M. Rahman, V. Govindaraju, and G. R. Thoma, "Biomedical article retrieval using multimodal features and image annotations in region-based CBIR," *Proceedings of the 17th Document Recognition and Retrieval Conference*, San Jose, CA, 2010.
23. B. J. Jansen and A. Spink, "How are we searching the world wide web?: a comparison of nine search engine transaction logs," *Information Processing and Management*, vol. 42, no. 1, pp. 248-263, 2006.

**Dina Demner-Fushman**

Dr. Dina Demner-Fushman is a staff scientist at the US National Library of Medicine. Her interest in biomedical language processing stems from years of clinical practice (M.D. obtained from Kazan State Medical Institute in 1980) and clinical research (Doctorate [Ph.D.] in Medical Science earned from Moscow Medical and Stomatological Institute in 1989.) She earned her M.S. and Ph.D. in Computer Science from the University of Maryland, College Park in 2003 and 2006, respectively. She earned her B.S degree in Computer Science from Hunter College, CUNY in 2000. Dr. Demner-Fushman is a fellow of the American College of Medical Informatics (ACMI).

**Sameer Antani**

Dr. Sameer Antani is a Staff Scientist at the Lister Hill National Center for Biomedical Communications, an R&D division of the US National Library of Medicine at the National Institutes of Health. He leads research in topics in multimodal biomedical information retrieval, content-based image retrieval, biomedical image processing, multimedia-rich interactive publications, mobile health apps, and a system to aid family reunification in mass disaster events. He earned his B.E. from the University of Pune, India, in Computer Engineering, and M.Eng. and Ph.D. from the Pennsylvania State University in Computer Science and Engineering. Dr. Antani is a member of SPIE, the IEEE and its Computer Society, and AMIA. He serves as the Vice Chair for Computational Medicine on the IEEE Technical Committee on Computational Life Sciences (TCCLS), and is the Chair-Elect for the AMIA Biomedical Imaging Informatics Working Group. He is an editorial board member of Elsevier Journal of Computers in Biology and Medicine.

## Matthew S. Simpson

Dr. Matthew S. Simpson is a postdoctoral fellow at the Communications Engineering Branch of the Lister Hill National Center for Biomedical Communications, a division of the US National Library of Medicine, National Institutes of Health. Dr. Simpson received his Ph.D. and M.S. degrees in electrical and computer engineering from the University of Maryland, College Park in 2011 and 2008, respectively, and his BS degree in computer engineering from Clemson University in 2004. Dr. Simpson's current research interests include biomedical image retrieval and natural language processing.

## George R. Thoma

Dr. George R. Thoma is Chief of the Communications Engineering Branch of the Lister Hill National Center for Biomedical Communications, a research and development division of the US National Library of Medicine at the National Institutes of Health. In this capacity, he directs intramural R&D in mission-critical projects, such as the automated extraction of bibliographic data from medical articles to populate the MEDLINE database, digital preservation, animated virtual books, a system to aid family reunification in mass disaster events, and multimedia-rich interactive publications. These projects rely on techniques from document image analysis, biomedical image processing and machine learning, and appear at archive.nlm.nih.gov. He earned his B.S. from Swarthmore College, and M.S. and Ph.D. from the University of Pennsylvania, all in Electrical Engineering. Dr. Thoma is a fellow of SPIE, the International Society for Optical Engineering.