Automatically Finding Images for Clinical Decision Support

Dina Demner-Fushman, Sameer Antani, George R. Thoma
Lister Hill National Center for Biomedical Communications
National Library of Medicine, National Institutes of Health, DHHS, Bethesda, MD 20894, USA
{ddemner,santani,gthoma}@mail.nih.gov

Abstract

Essential information is often conveyed in illustrations in biomedical publications. A clinician's decision to access the full text when searching for evidence in support of clinical decision is frequently based solely on a short bibliographic reference. We seek to automatically augment these references with images from the article that may assist in finding evidence.

The feasibility of automatically classifying images by usefulness (utility) in finding evidence was explored using supervised machine learning. We selected 2004 -- 2005 issues of the British Journal of Oral and Maxillofacial Surgery, manually annotating 743 images by utility and modality (radiological, photo, etc.) Image data, figure captions, and paragraphs surrounding figure discussions in text were used in classification.

Automatic image classification achieved 84.3% accuracy using image captions for modality and 76.6% accuracy combining captions and image data for utility.

Our results indicate that automatic augmentation of bibliographic references with relevant images is feasible.

1. Introduction

Clinicians can fairly accurately form an opinion about the relevance of a publication to a clinical situation based on its title alone; however the title is not always sufficient in determining the Evidence-Based Practice (EBP) usefulness (henceforth evidence-based utility or clinical utility) of a publication [1].

Given that medical illustrations often convey essential information in compact form, we seek to automatically identify illustrations that could help clinicians evaluate the potential usefulness of a publication in a clinical situation. We hypothesize that in many cases a short outcome statement that is currently automatically extracted by Demner-Fushman et al to augment the title of a MEDLINE citation [2] could be rendered more useful if accompanied by one or more extracted images. For example, given a title "Clinical management and microscopic characterization of fatigue-induced failure of a dental implant" a clinician might not know if the article applies to the case for which evidence is sought. The automatically extracted outcome statement in Figure 1 will clarify that the "fatigue-induced failure" is a fracture, and the adjacent x-ray will illustrate what is meant by the "typical signs" of a fracture. Before testing the hypothesis about the value added by images, however, we need to establish if automatic image annotation by utility for EBP is attainable, and if such images can be reliably extracted from the original articles.

The fractured implant showed the typical signs of a fatigue-induced fracture in the coronal portion of the implant together with numerous micro-fractures in the apical one …

Figure 1. A fractured implant and relevant fragment of the automatically extracted outcome statement (reproduced with author's permission [3])

The long-term goals of the project are to develop robust algorithms for automatic biomedical image annotation by utility for EBP. The current study explores the feasibility of such annotation. The novelty of this work is in proposing and developing a classification of images by their usefulness to support EBP.

2. Background

The importance of medical illustrations in clinical decision making has motivated the development of large databases of medical images, such as the Public

Health Image Library (PHIL)[1], as well as active research in image retrieval [4]. However, many systems continue to implement image retrieval based on manually generated textual descriptions of images, such as at Isabel Healthcare[2]. The high cost of such labor-intensive annotation has spurred research in automatic image annotation [5] by content-based image retrieval (CBIR) techniques [6]. Advances in biomedical image retrieval from large databases have been evaluated since 2004 in the yearly ImageCLEFmed tasks of the Cross Language Evaluation Forum (CLEF) [7]. In the 2006 ImageCLEFmed evaluation the best performing systems achieved 16-17% error rate in the image annotation task using supervised machine learning techniques based on combined image and textual features [8].

The successes in image annotation, CBIR, and text classification based on image captions motivated integration of image data for biomedical text categorization [9]. Other efforts explored biomedical article retrieval based on image content [10,11], and use of textual and image features for image classification in biomedical articles [12]. While preliminary studies on image-only retrieval have resulted in mediocre results, classification of bioscience images into six generic categories achieved an average F-score of 73.66% [12].

Encouraged by the success achieved in various informatics applications through combining textual and image data, our study explores a new area of biomedical image annotation using textual and image data – that of classifying images in biomedical articles with respect to their utility for clinical decision support.

3. Methods

We selected the *British Journal of Oral and Maxillofacial Surgery*[3] because it is the area of expertise of the first author and it is representative of the specialties that might particularly benefit from visual information presented early in the information retrieval process. Two authors of this paper studied 2004 and 2005 online issues of the journal and downloaded full-size images and HTML text of articles and short communications containing images. The images were then manually annotated and cross validated for their modality and utility in finding clinical evidence. Supervised machine learning techniques were applied to evaluate feasibility of automatic annotation of images by clinical utility.

3.1 Manual Image Annotation

We established three facets of EBP essential in finding and automatically extracting textual information for clinical decision support: 1) knowing the clinical task; 2) identifying elements of a clinical scenario (patient/problem, intervention/comparison, and outcome); and 3) determining the strength of evidence of the medical article [2]. In a preliminary analysis of the 2006 issues of the journal, we found that of these facets the elements of the clinical scenario for the clinical tasks of diagnosis and therapy are well illustrated. Two image modalities that provide valuable additional information to a clinician without requiring a significant cognitive effort are photographs and radiological images.

Table 1. Image modality categories (NN – number of images in a given category)

Category	Definition	NN
Chart/ Graph	A geometric diagram consisting of dots, lines, and bars.	108
Drawing	A hand drawn illustration	70
Flowchart	A symbolic representation of sequence of activities.	6
Form	A compilation of textual data and/or drawings related to patient and/or clinical process.	10
Histology	An image of cells and tissue on the microscopic level.	134
Photograph	Picture obtained from a camera	252
Radiology	A 2D view of an internal organ or structure (includes X-ray, CT, PET, MRI, ultrasound)	101
Table	Data arranged in a grid	47
Mixed	Images combining modalities (e.g., drawings over an x-ray)	15

It was not clear a priori if determining the clinical utility of an image requires the image modality to be known. Therefore, we developed two annotation schemes: 1) by image modality, and 2) by evidence-based utility. To create a reference standard set we manually annotated all extracted images with categories from both schemes Table 1 presents the first scheme that takes into consideration attempts to reconcile previously identified categories of figures in scientific documents [8,9,12,13].

[1] http://phil.cdc.gov/Phil/home.asp
[2] http://www.isabel.org.uk
[3] http://intl.elsevierhealth.com/journals/bjom

Table 2 presents the evidence-based utility classification of images. Both tables present the number of images in the reference standard set assigned strictly to a given category. At present, we classify all modality categories except for *Photograph* and *Radiology* images as *Other* in terms of their evidence-based utility. In future work we plan to classify other modalities to a higher degree of granularity.

We noticed that figures containing tables can be easily identified using captions (e.g., "Table 1…"). Therefore tables were not used in the annotation experiments. We also disregarded flowcharts because there were few such figures in our sample. For the same reason we did not use 15 images with various mixed modality annotations and 10 images presenting multiple utility categories. Thus 675 of 743 extracted images were selected for the modality and utility annotation experiments.

Table 2. Image utility for Evidence-Based Practice

Category	Definition	NN
Diagnostic (Dx)	Image presenting distinct characteristics of a disorder (Supports diagnostic task; presents patient/problem)	145
Instrument/ Artifact (Ix)	Image of medical instrument, device, or an artificial substitute for a missing body part (Supports diagnostic and therapy tasks; presents intervention)	39
Procedural (Px)	Image presenting details and/or steps of a clinical intervention (Supports diagnostic and therapy tasks; presents intervention)	68
Result, Evidence (Ex)	Image presenting the results of intervention or a disease (Presents outcome)	81
Mixed (Mx)	Image presenting several utility categories (e.g., a result of an intervention and an instrument)	15
Other (Ox) (excluding tables and flowcharts)	Images not directly pertaining to a clinical situation, e.g., portraits, computer graphics, or a result of gel electrophoresis.	342

3.2 Text Pre-processing

The HTML formatting of the articles in the journal is relatively uniform and well-structured, which allowed use of regular expressions to extract figure captions and the lines of text surrounding each discussion of a figure in a paragraph. Additional patterns were developed to identify parts of captions corresponding to panels of the multi-panel images described above. For example, references to image panels (A) and (B) are extracted from the caption: "Figure 1. (A) Intraoral view before … (B) Intraoral view two years postoperatively …" In case of an error in panel identification, the extracted caption and paragraph text was duplicated for each panel.

3.3 Image Pre-processing

Images in the articles were in one of two formats: GIF and JPEG. The variations in image file format, figure file naming convention, and number of color channels in the image required manual extraction and renaming of the image files. Additionally, there were several multi-panel images that were referred through subfigure labels in the captions, but were stored as a single image file in the electronic article. These images required manual cropping and labeling. The file format and color channel problem were addressed through automatic image normalization techniques applied before computing image features for automatic image annotation.

These preprocessing steps yielded a set of single-paneled images each accompanied by the automatically extracted caption and text paragraphs. The images are uniquely identified by a name that encodes information about the source article, figure number and, if applicable, panel/subfigure reference. As future work, we plan to automate extraction and labeling from analysis of HTML text, and cropping of sub-images from a multi-panel image.

3.4 Automatic Image Annotation

Image and textual data obtained above were explored using YALE[4], a freely available open source machine learning environment. We used its Word Vector Tool plugin to represent the extracted text as feature vectors needed for machine learning. In addition, we experimented with replacing extracted text with preferred names of the UMLS concepts identified in the text using MetaMap [14].

The image feature vectors were obtained using methods developed in MATLAB[5]. As a preliminary

[4] http://rapid-i.com
[5] http://www.mathworks.com

approach, texture and color features were computed on the entire image without applying any image segmentation techniques. Texture features were computed as a 3-level discrete 2-D Daubechies' wavelet transform [15]. Though several color features were evaluated, the four most dominant colors and their extent computed in the perceptually uniform CIE LUV color space [16] proved most effective.

Preliminary experiments confirmed that, as for many other text and image multi-class classification problems, SVM type 1[6] with radial basis function (RBF) kernel, cost parameter C=1,000, and the RBF width parameter $\gamma = 0.01$ performed well for both our annotation tasks [17]. The libSVMLearner with these settings was used to conduct evaluations using the following features and their combinations: 1) captions; 2) text surrounding figure discussions; 3) UMLS concepts; 4) image data; 5) captions + discussions; 6) captions + Image data

Table 3. Image annotation accuracy (in %). (Image-t refers to the texture measure and Image-d to dominant colors.)

Features	Modality	Clinical Utility		
		direct	photo	x-ray
Individual features				
Caption	**84.3 ±2.6**	75.6 ±5.4	72.5 ±16.3	71.4 ±10.1
Discussion	77.4 ±6.8	67.1 ±4.3	53.7 ± 8.1	76.3 ± 7.6
UMLS	79.6 ±5.3	70.4 ±2.7	55.3 ± 8.3	64.0 ±9.1
Image-t	79.4 ±3.1	69.3 ±2.5	46.7 ± 7.9	64.0 ±10.2
Image-d	67.3 ±5.1	63.6 ±3.0	34.3 ± 5.0	63.0 ± 9.0
Combined features: C=caption, D = discussion, I = image (texture + color)				
I	80.6 ±3.5	70.7 ±4.0	45.0 ±10.1	63.0 ± 6.4
C + D	82.8 ±4.3	76.4 ±3.9	63.3 ± 8.7	78.2 ± 9.9
C + I	84.0 ±3.6	**76.6 ±4.2**	65.4 ± 4.6	70.0 ± 8.9

For fusion of textual and image data, features were combined by joining data sets generated for individual feature sources. These evaluations were conducted for modality and utility annotation. Clinical utility annotation was tested under two conditions: 1) within a single modality for photos and radiology images and 2) on the whole dataset directly for utility, disregarding the modality. To avoid propagating automatic modality annotation errors, we established an upper bound for utility classification within a single modality by using the manually assigned modality classes. We evaluated the accuracy (relative number of correctly classified images) of automatic image annotation using YALE 10-fold cross validation procedure.

[6] http://www.statsoft.com/textbook/stsvm.html

Table 4. Confusion Matrix for utility classification based on joint caption and image data (C+I) compared to captions alone (C). (Number of correctly identified images are shown in bold.)

True (Total)	Feature	Dx	Px	Ex	Ix	Ox + Mx
Dx (145)	C+I	**101**	9	23	15	15
	C	**104**	6	28	2	6
Px (68)	C+I	5	**40**	8	9	2
	C	2	**38**	9	8	1
Ex (81)	C+I	16	5	**42**	1	3
	C	5	2	**32**	0	1
Ix (39)	C+I	1	0	1	**6**	0
	C	1	0	0	**9**	2
Ox + Mx (357)	C+I	22	14	7	8	**322**
	C	33	22	12	20	**332**

4. Results

The accuracy of modality annotation based on captions and image data (combined texture and color features) is comparable to results reported in the literature [8,12]. Our utility annotation results are less accurate than modality annotation, but still quite encouraging. Table 3 presents annotation results for all our experiments.

The results of combining captions and image data for modality and utility classification do not differ much from the results based on captions alone. Significance tests will be conducted upon completion of tests on a larger collection spanning multiple journal types and years. Most misclassifications occurred for results classified as diagnostic procedures (See Table 4.) Adding image features to captions reduced the number of misclassifications into the *Other* category, but significantly increased the number of images in the *Other* category categorized as *Diagnostic*. Image features increased correct classification into *Procedural* and *Results* categories.

5. Discussion

Overall the results of automatic image annotation by their evidence-based utility are very encouraging and consistent with the state-of-the-art in text and image classification. While it may seem that the gains are insignificant, if any, in combing text and image modality, an improvement can be seen in applying the proposed technique for direct clinical utility in support of EBP.

There were several unexpected findings in this study. The first being a wide range of image modalities

and content in this highly specialized journal (such as drawings of leeches in a historical article, or a photograph of electrophoresis of products of polymerase chain reaction for analysis of immunoglobulin gene rearrangement. Another somewhat surprising finding is the fact that knowing an image modality does not help in determining its clinical utility. Exceptions to this might be histology images which can be identified with high accuracy through texture measures, and charts and graph images by virtue of their color distribution which tends to use few and usually uniform colors.

Another unexpected result is the slight deterioration of classification results when the figure discussions in article text are included with figure captions. One explanation for this phenomenon might be that although the discussions provide more details about a given procedure or result, the utility of an image is more precisely defined in the caption. For example, given a caption "A non-ulcerated soft-tissue swelling in the left lower lip" one might assume that the image has diagnostic utility, which is correct. Given the discussion text for the same figure, *"He had a 1cm slightly bluish, soft-to-firm, non-tender circumscribed swelling on the left side of the lower lip. We diagnosed a mucocele and excised the lesion under local anaesthesia"*, however, it is hard to guess whether the image is diagnostic, or that of the results of the excision. The fairly large misclassification of the *Results* as *Diagnostic* images and conversely could be explained by the fact that the images often present the same patient before and after treatment. We plan to add information about an image position in a sequence to the features for machine learning. This might help in distinguishing between the *Diagnostic* and the *Results* images.

7. Conclusions and Future Work

Although our findings are sufficient for establishing feasibility of image annotation by clinical utility, there are limitations to our study. As we studied images from only one journal, our findings need to be confirmed on a variety of clinical journals. The number of images in issues published over two years was sufficient for cross-validation, but the range of modality classes and sparseness of the diagnostic, procedural, and results images did not allow for dividing the collection into training, testing, and validation sets. Our collection needs to be expanded by adding more journals and covering larger time spans for each.

This study demonstrates that images presented in clinical journals can be successfully annotated by their usefulness in fining evidence to assist a clinical decision. Using methods that achieve state-of-the-art results (84.3% accuracy) in modality annotation, we achieve 76.6% accuracy in clinical utility annotation.

The feasibility of automatic image classification with respect to its utility in finding clinical decision support demonstrated in this study provides several venues for further exploration. We plan to study the influence of augmenting bibliographic references retrieved from a database search with images; new ways of organizing and presenting retrieval results using annotated images; and further improvement in the automatic single and multi-panel image extraction, annotation, and complementary text extraction.

Acknowledgment

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 Communication (LHNCBC).

References

1. D. Demner-Fushman, S. Hauser, G. Thoma. The role of title, metadata and abstract in identifying clinically relevant journal articles. AMIA Annu Symp Proc; 2005:191-5.
2. D. Demner-Fushman, J. Lin. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics. 2007;33(1):63-104.
3. S. Capodiferro, G. Favia, M. Scivetti, G. De Frenza, R. Grassi. Clinical management and microscopic characterisation of fatigue-induced failure of a dental implant. Case report. Head Face Med. 2006 Jun 22;2:18.
4. H. Müller, H. Michoux, D. Bandon, A. Geissbuhler. A review of content-based image retrieval systems in medicine - clinical benefits and future directions. International Journal of Medical Informatics, 2004; 73:1-23.
5. T.M. Lehmann, M.O. Güld, C. Thies, B. Plodowski, D. Keysers, B. Ott, H. Schubert. IRMA – Content-based image retrieval in medical applications. MEDINFO 2004:842-8.
6. S. Antani, L.R. Long, G. Thoma. Content-based image retrieval for large biomedical image archives MEDINFO 2004:829-33
7. W.R. Hersh, H. Muller, J.R. Jensen, J. Yang, P.N. Gorman, P. Ruch. Advancing biomedical image retrieval: development and analysis of a test collection. J Am Med Inform Assoc. 2006 Sep-Oct;13(5):488-96.
8. H. Müller, T. Deselaers, T.M. Lehmann, P. Clough, E. Kim, W.R. Hersh. Overview of the ImageCLEFmed

2006 medical retrieval and annotation tasks. Working Notes for the CLEF Workshop. 2006 20-22 Sep.Alicante, Spain.
9. H. Shatkay, N. Chen, D. Blostein. Integrating image data into biomedical text categorization. Bioinformatics. 2006 Jul 15;22(14):e446-53.
10. T.M. Deserno, S. Antani, L.R. Long. Exploring access to literature using content-based image retrieval. SPIE Medical Imaging 2007. vol. 6516.
11. A. Christiansen, D.-J. Lee, Y. Chang. Finding relevant PDF medical journal articles by the content of their figures. SPIE Medical Imaging 2007. Vol. 6516
12. B. Rafkind, M. Lee, S.F. Chang, H. Yu. Exploring text and image features to classify images in bioscience literature. Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology. 2006 Jun:73-80.
13. X. Lu, P. Mitra, J.Z. Wang, C.L. Giles. Automatic categorization of figures in scientific documents. Proceedings of the Joint ACM/IEEE Conference on Digital Libraries, Chapel Hill, North Carolina. 2006 Jun:129-138.
14. A.R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.
15. R.C. Gonzales, Woods RE. Digital Image Processing 2nd Ed. Prentice Hall, NJ, USA.
16. Y. Deng, B.S. Manjunath, C. Kenney, M.S. Moore, H. Shin. An Efficient Color representation for image retrieval. IEEE T Image Proc. 2001 Jan; 10(1):140-7.
17. R.E. Fan, P.H. Chen, C.J. Lin. Working set selection using second order information for training SVM. Journal of Machine Learning Research 2005;6:1889-1918.