Does Figure-Text Improve Biomedical Article Retrieval?

Daekeun You, Sameer Antani, Dina Demner-Fushman, and George R. Thoma
Lister Hill National Center for Biomedical Communications / NLM

1. INTRODUCTION

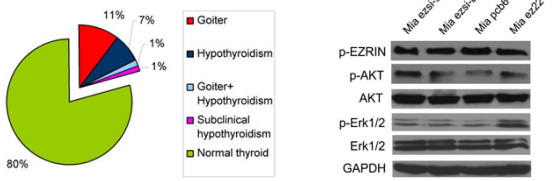
Figures with graphical illustrations (e.g., graphs, charts, diagrams) are often used in biomedical articles to convey statistical results, biomedical procedures, schematics, etc. They are frequently accompanied with superimposed graphical text annotations, or "figure-text". Recently researchers reported that about 70% of figure-text occurring in illustration figures are not found in the associated figure captions, and implying that extracting it added value to the indexes. The result is interesting because it has been assumed that:

- ◆ figure-text provides information that complements associated textual metadata, such as the captions, or bibliographic citations, and
- ◆ using figure-text for indexing figures enhances the quality of information retrieval.

However, we find nothing in the literature that adequately supports this assumption of information gain due to figure-text. In other words, use of figure-text for biomedical image retrieval is considered important for correlating text metadata, such as figure captions and mentions, with the visual material. However, this does not necessarily mean that the additional information is actually useful in improving image retrieval or for answering clinically meaningful questions. To our knowledge, no convincing evaluation results have been reported that address this uncertainty.

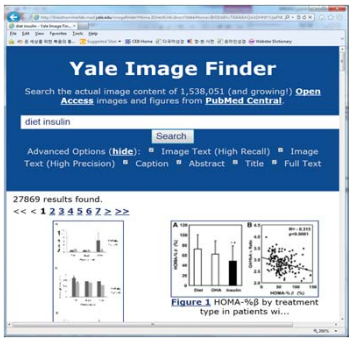
In our research, we attempt to ascertain the importance of figure-text in indexing figures in research articles by evaluating the information gained. Our goal is to ***evaluate the advantage offered by indexing figure-text in addition to conventional text metadata*** such as figure caption, mentions, and article text for biomedical image indexing and retrieval. To this end, following steps were done:

- ◆ manually extract figure-text from images in our dataset,
- ◆ create three document sets that consist of figure caption alone, figure caption and figure-text in combination, and figure-text alone,
- ◆ index the image-associated text using Apache Lucene
- ◆ conduct several retrieval experiments and compare the results.



Samples of illustration figures containing figure-text

2. PRIOR WORK



YIF search engine utilizing figure-text for images retrieval

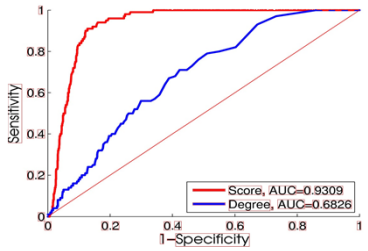


Figure-text detection from a graph figure

3. METHOD

Dataset creation: Selected "**lung cancer**" as our topic for two reasons.

- ◆ it is a fairly broad topic that has many sub-topics (e.g., symptoms, causes, diagnosis, and treatment) allowing for the creation of interesting and meaningful queries (questions for image retrieval), and
- ◆ it can provide a sufficiently large number of images that may answer the questions.

We submitted the search term "lung cancer" to the **OpenI** system (**http://openi.nlm.nih.gov**) and obtained 2,006 images as a result. We split the set into two and used one set (containing 806 images) for pilot system development and retrieval test.

Figure-text extraction: We manually extracted figure-text from the dataset by cropping figure-text regions from figures, recognizing the text regions using Tesseract (an open source OCR engine), and then correct the OCR results.

Figure-text extraction results

Number of total extracted figure-text words	13,413
Number of total unique figure-text words	12,512
Number of images with new figure-text	717
Number of total new figure-text words	6,016
Number of gene/protein/cell line figure-text words	2,506
% of total new figure-text	48.1%

Document indexing and retrieval: An image is represented by three different types of documents (text data) for indexing and retrieval purpose.

- ◆ **cap-doc:** figure-caption alone
- ◆ **ext-doc:** figure-caption + figure-text (a combined text data)
- ◆ **FT-doc:** figure-text alone

The open-source **Apache Lucene** search engine library was used to implement an indexing and search system of the images using the text data (documents). The search system enables image search by text queries and return a ranked-list of relevant documents that can be used to reference the images.

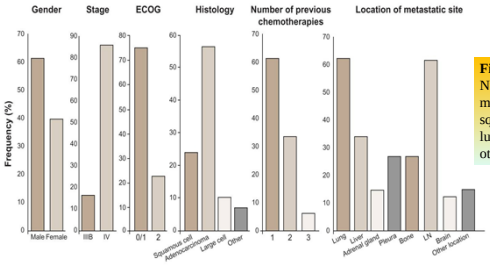


Figure-text: Gender Stage ECOG Histology Number of previous chemotherapies Location of metastatic site frequency male female IIIB IV squamous cell adenocarcinoma large cell other lung liver adrenal gland pleura bone LN brain other location

Figure caption: Demographic distribution of population from a Phase II randomized double-blind study with BIBF 1120 as monotherapy in advanced non-small cell lung cancer. Abbreviations: ECOG, Eastern Cooperative Oncology Group score; LN, lymph node.

An illustration figure and its figure caption (*cap-doc*) and figure-text (*FT-doc*)

4. EVALUATION & RESULT

Evaluation strategy

- ◆ **Query creation (on "lung cancer"):** Two expert judges created 10 questions;
- ◆ **Retrieval test:** (i) submit the 10 queries to the pilot search system, (ii) choose top-5 retrieved documents for each query from each dataset, and (iii) pose the selected documents to the judges for relevance judgment.
- ◆ **Relevance judgment:** Experts marked system responses as "Relevant", "Maybe" (partially relevant), "Irrelevant"

10 questions on "lung cancer" and retrieval results

Q1	Does the chance of getting lung cancer increase with age?
Q2	How does the lung cancer mortality rate compare to the mortality rate for other cancers?
Q3	How many people die in the US from lung cancer each year?
Q4	What gene mutation cause lung cancer?
Q5	What genes are involved in lung cancer?
Q6	What is the 5-year survival rate of people with lung cancer?
Q7	What is the increased risk of lung cancer for those individuals who smoke?
Q8	What is the risk of developing lung cancer for patient with COPD?
Q9	What percentage of newly diagnosed patients undergo chemotherapy for lung cancer?
Q10	What primary cancers can metastasize to the lung?

Questions	*cap-doc*		*ext-doc*			Figure-text used for judgment	=	*FT-doc*	
	Ret	R-R	Ret	R-R	New R-R			Ret	R-R
Q1	5	1	5	1	0	4/6	5	2	0
Q2	3	2	4	2	0	4/5	3	2	2
Q3	5	1	5	3	2	4/7	3	3	3
Q4	5	3	5	3	0	3/10	5	5	4
Q5	5	4	5	3	1	8/10	3	5	4
Q6	5	3	5	2	0	6/6	4	2	2
Q7	5	2	5	3	1	6/9	4	5	4
Q8	5	2	5	2	0	4/6	4	5	1
Q9	5	1	5	1	0	5/8	4	5	1
Q10	5	1	5	0	0	2/12	1	5	0

Ret: **R**etrieved
R-R: **R**elevant and **R**etrieved
New R-R: Number of relevant documents **exclusively retrieved by *ext-doc*** compared to *cap-doc*

5. CONCLUSION

Figure-text could be useful for

- ◆ searching images with specific members (terms) within a category (e.g., "genes", "cells", "countries"). Figure captions generally mention the categories but often do no enumerate all members
- ◆ *query by image* search: figure-text may be the only source that can provide essential information about the content that is conveyed in the query image

Figure-text may help improve retrieval quality for a limited number of specific queries. However, using figure-text with other text metadata (e.g., figure caption) may not significantly change the retrieval rank. Also, a number of documents can only be retrieved by figure-text: however, they may generally rank low when figure-text is mixed with other metadata.