U.P.B. Sci. Bull., Series C, Vol. 70, No. 2, 2008 ISSN 1454-234x

EXTRACTION AND INTERPRETATION OF IMAGE-RELATED MEDICAL KNOWLEDGE FROM ON-LINE HEALTH-DOCUMENTS

Filip FLOREA[1], Vasile BUZULOIU[2], Alexandrina ROGOZAN[3], Valeriu CORNEA[4], Abdelaziz BENSRHAIR[5], Srefan DARMONI[6]

În ultimii 10 ani, Internetul a devenit o sursă primară de informaţie, în multe domenii, printre care şi cel al sănătăţii. Catalogul medical Francez CISMeF a fost conceput pentru a oferi funcţionalităţi de indexare şi căutare a documentelor medicale. O mare parte din informaţia medicală disponibilă on-line este prezentă sub formă de imagini, acestea fiind un important instrument de diagnostic şi învăţământ. Pentru a imbogaţii motorul de căutare Doc'CISMeF cu o funcţie de "căutare-după-criterii-imagistice", grupul nostru a dezvoltat modulul MedIC, cu scopul de a permite utilizatorilor să caute documente folosind criterii imagistice. Modulul MedIC extrage descrieri ale imaginilor luând în considerare atât imaginea propiu-zisă, cat şi regiunile textuale corespunzatoare imaginii. În această lucrare, descriem algoritmul de extragere şi interpretare a regiunilor textuale, precum şi rezultatele obţinute pentru extragerea modalitaţii medicale a imaginilor.

Over the last decade, Internet has become a significant source of information in numerous fields, including health. The French heath-catalogue CISMeF provides medical documents indexing and searching capabilities. The images represent a significant portion of the on-line medical knowledge and a valuable component of diagnosis and teaching. To enrich the Doc'CISMeF search engine with a "search-by-image-criteria" capability, we developed the MedIC module to allow the users to search for documents using image-related criteria. To extract accurate image descriptors, the MedIC architecture takes into consideration both the image itself and the corresponding image-related text-regions. In this paper we present our approach for text-region extraction and interpretation, as well as the results obtained for image medical modality extraction.

Keywords: Internet, medical information extraction, image annotation.

[1] PhD Student, Image Processing and Analysis Laboratory, University POLITEHNICA of Bucharest, Romania
[2] Prof., Image Processing and Analysis Laboratory, University POLITEHNICA of Bucharest, Romania
[3] Reader, LITIS Laboratory, INSA de Rouen, France
[4] Student Master, INSA de Rouen, France
[5] Prof., LITIS Laboratory, INSA de Rouen, France
[6] Prof., LITIS Laboratory, University of Rouen, France

1. Introduction

The development of the Internet, as well as the advances in the way the data is acquired, stored and shared, made the health information freely available in ever growing quantities.

CISMeF (French acronym for Catalog and Index of French-language health resources) is a quality-controlled subject gateway [1], initiated by the Rouen University Hospital in 1995 [2]. Its role is to provide online searching capabilities for health resources (*i.e.* medical documents), by describing and indexing the most important documents of institutional health information in French. From the beginning CISMeF is indexed manually by a team of experienced health-librarians. It currently contains more than 16,500 resources (at 1st December 2006), and it is updated manually with 50 new resources each week. Considerable efforts were undertake by the CISMeF team to develop automatic textual indexation architectures, and recently significant advancements were reported [3]. However, the automated indexing has drawbacks, and one of the biggest is the difficulty of indexing non-textual media, like *the images*.

Medical imaging has grown over the last decade to become an essential component of diagnosis, medical teaching and civic education. The development of Internet technologies has made medical images available in large numbers in online repositories, collections, atlases, and other heath-related resources. These images are representing a valuable source of knowledge and are of significant importance for medical information retrieval. However, the shear amount of medical visual data available online makes it very difficult for users to locate exactly the images that they are searching for. The annotation of images with relevant image-related medical keywords could allow the use of textual queries to search medical images. However, the cost of manually annotating images is prohibitively high as it is time-consuming and requires medical knowledge. To provide efficient and fast access to medical visual-data, automatic systems for extracting relevant medical information from images are needed.

2. Paper contents

In the context of automatic health-resource indexing, the CISMeF team developed the MedIC (Medical Image Categorization) module, to automatically extract medical information from the images included in on-line documents. The goal of MedIC is to add *search-by-image-criteria* capabilities to the Doc'CISMeF search engine, and thus, to provide the users (*e.g.* health professionals, students or general public) with the possibility of using *image-related terms* when performing queries.

The MedIC architecture was designed to extract several types of image-related medical information:

‹1› medical modality (*e.g.* standard radiography, magnetic resonance),
‹2› anatomical region, biological system and/or organ,
‹3› acquisition view-angle (*e.g.* axial, coronal, sagittal),
‹4› pathology/disease,
‹5› other acquisition parameters (*e.g.* contrast agents, specific measures);

at several levels:

‹a› carried by the visual content of images,
‹b› annotated directly on the image,
‹c› carried by the image-related text-regions of the document (*e.g.* image caption, image name, the paragraph/sentence that points to the caption number).

The semantic information provided by the MedIC module ‹1›→‹5› is intended to be added to the index of each document, to allow users to formulate queries containing *image-related keywords* (*e.g.* "find me all the resources containing **lung computed tomography** images") in addition to the *document-related keywords* currently used (*e.g.* "find me all the resources related to **pulmonary embolism**").

Encouraging results were already obtained when using MedIC to extract medical information from the image it-self: from the image visual content ‹a› or from the annotations "marked" directly on the image ‹b›. The first approach is based on the description of images using texture and statistical features, and uses machine learning for the categorization of the resulted numerical image representation in pre-defined classes. Accuracies of up to 96% were noted for the recognition of ‹1›, ‹2› and ‹3›, and +98% when only ‹1› was needed [4]. In a second experiment we used MedIC to extract and interpret the text directly "marked" on the images ‹b›. We use image morphology to extract the text from images, and Optical Character Recognition and modality-related production rules defined by radiologists, to interpret the text and extract the modality. This way, once again, modality recognition accuracies of more than 98% were noted [5].

Given that the medical documents are carrying textual information related to the images they carry (this information is contained in image captions, image names and image-related paragraphs), we designed and implemented a third approach aimed at the automatic extraction and interpretation of the image-related text-regions ‹c›. In this paper we are introducing this approach and we are

presenting the results of a series of experiments performed to evaluate the significance of the information carried by the text-regions.

The rest of this paper is organized as follows. Section 2 describes some of the related work. In Section 3 we are introducing the extraction of the document image-related text-regions and in section 4 we are continuing with the interpretation of their content by defining medical dictionaries and extracting relevant terms for each type of information we are searching for (*i.e.* ‹1›→‹5›). We then present, in section 5, experimental results obtained for modality recognition, on a set of 718 image-texts pairs extracted from documents indexed by CISMeF. We conclude this paper by presenting discussions, conclusions and perspectives in Section 6.

3. Related works

There are a number of systems that are treating medical images in the form of medical image categorization, indexing and/or retrieval applications. These systems are using numerical representation of the visual content to either classify the images in predefined medical categories or to retrieve the most similar images (from a previously indexed image database) when presented with a new image. The majority of the existing systems are related to specific modalities: KMeD [6] and COBRA [7] are treating MRI head images, ASSERT-system deals with lung CT images [8], I-Browse [9] operates on histological slices and [10][11] with X-Rays. Given the fact that the principles used by each of these systems are highly-dependent of the particular conditions of each medical modality, they are not directly and/or entirely applicable to other cases. However systems better adapted to cope with various image modalities, anatomical regions and pathologies were proposed more recently: MedGIFT [12] and IRMA [13].

The visual content of medical image is an extremely rich source of medical knowledge, but capturing and interpreting all this information has proved to be a significantly complex and difficult task. Therefore, as we said, the MedIC system is designed to extract information not only from the content of medical images ‹a› [4], but also from the annotations "marked" directly on the image (when these annotations exist) ‹b› [5] and from the image related text-regions of the containing document ‹c›. This is possible because the images are extracted directly from on-line medical resources, enabling us to extract a series of image-related text-regions (image captions, image -related paragraphs) that are presenting/discussing the content of the images.

To our knowledge this type of multi-source, image-text approach was never proposed for automatic extraction of medical image-related information (categorization, indexing and/or retrieval of medical images) and none of the medical information search engines freely available on the Internet use it.

4. Extraction of the document image-related text-regions

In our context of application, the medical images are part of heath-resources (*i.e.* mainly HTML – 68.10% and PDF- 25.20% documents). Contrary to the case of images extracted from medical repositories, important information about the images nature can be extracted from the resource (*i.e.* document). As the image is an illustration of a fact, experiment or observation presented in the resource, there is usually at least one phrase or paragraph explaining its content. Furthermore, in the majority of the resources following an academic format (*i.e.* published papers or books, technical reports - usually in PS, PDF, DOC or RTF formats) the images have captions (usually one or two sentences, placed near the image) summarizing and explaining the information presented by the image. Using this caption, a link between the image and the image-related paragraph can be obtained (*e.g.* "as shown in Fig. 3"). Preliminary studies pointed out considerable difficulties in correctly map the image with the corresponding caption (*e.g.* several images located in the proximity of only one caption) or the corresponding paragraph (*e.g.* instead of clear referencing like "see Fig. 3", forms like "the above figure" or "the figure seen at page 4" are sometimes used). Furthermore, the mapping between the images and image-related text paragraphs was often impossible because of the absence of either the image caption, the image numbering (*e.g.* "Fig. 3") or the image referencing by its number (*e.g.* "see Fig. 3").

The way the images, captions and paragraphs are related, formatted and referenced is highly dependent on the document-format. Contrary to academic publications, the Internet hypertext documents have much less restricted formats. Therefore it is more frequent to find images with no captions, or several images that are sharing a single caption. The fact that HTML is not restricted to a fixed page-width, like the printed papers are, allows the placement of many images on a single line. Therefore even when a caption is present, it is difficult to automatically determine which image/s is/are related to which part of the caption. However, this also imposes fewer restrictions on the placement of objects in documents, the hypertext objects being positioned one after the other (whereas for the fixed-width documents, the image objects are inserted in the first available space). This proximity of the objects is very significant because this way it is easier to map the images with their related text-regions (usually the object that precedes/follows the image). All this diversity and variability calls for different approaches when dealing with documents following printable (*e.g.* PDF and PS, but also DOC and RTF) and hypertext format (*i.e.* HTML). Furthermore, a number of these document formats are closed/proprietary (especially PDF – Adobe Systems® and DOC, PPT/PPS - Microsoft®), which makes the automatic extraction of the image and text objects even more difficult.

Ongoing experiments are aimed at the document structure understanding and image extraction from HTML (*i.e.* the most common format on CISMeF) and PDF (*i.e.* the common format for exchanging printable documents on Internet) using several Java open-source decoding and access libraries: PJ/PJX (http://sourceforge.net/projects/pjx) and PDFBox (http://www.pdfbox.org/) for PDF and HTMLParser (http://htmlparser.sourceforge.net/) for HTML.

Once the text-regions related to a given image is extracted we proceed to the interpretation of their content using a methodology derived from the automatic text indexing approach that the CISMeF team is developing [3].

5. Interpretation of the text regions

For the experiments presented in this paper we chose to test this approach on the extraction of the image medical modality. A medical modality represents any of the various types of equipment or probes used to acquire images of the body (*e.g.* radiography, ultrasound, computed tomography). Each of the acquisition modalities used in modern medicine is based on different physical principles (*e.g.* X ray penetration, magnetic resonance), and therefore, these modalities are used to capture particular anatomical, biological and/or pathological characteristics. Consequently the information present in medical images is highly dependent on the acquisition modality. This makes the modality, not only the most general image acquisition characteristic, but also the first image information to be extracted for medical information retrieval.

For testing this approach we extracted from web-resources indexed by CISMeF a set of 718 images along with their related text-regions. The "image-texts" pairs are extracted from both HTML and PDF documents indexed in CISMeF. The images are representing the main six categories of medical-imaging modalities: Standard Angiography (Angio), Ultrasonography (US), Magnetic Resonance Imaging (MRI), Standard Radiography (RX), Computed Tomography (CT), and Nuclear Scintigraphy (Scinti). The repartition of the images in the six modality categories as well as the number and proportion of images for which related text is available (images with text) are presented in Table 1.

The documents are covering various topics, from didactic materials to clinical cases and patient files. Therefore there is significant variability in the way the information is presented (*i.e.* from basic medical principles to advanced diagnosis methods and pathologic treatment) or what the images are illustrating (multiple modalities, anatomical regions, biological systems, acquisition parameters and pathologies).

Due to the absence of relations between some images and their corresponding texts or the difficulties we encountered with the correct extraction of this correspondence, only 585 entries of our database have complete "image-

texts" pairs (images for which related text is available and correctly extracted). For these cases we proceed to extract relevant image modality information using the text.

Table 1

Dataset composition

		№ of images		№ of images with text	
		absolute	relative	absolute	relative
Standard Angiography	Angio	104	14.48%	65	11.11%
Ultrasonography	US	40	5.57%	37	6.32%
Magnetic Resonance Imaging	MRI	88	12.26%	88	15.04%
Standard Radiography	RX	314	43.73%	243	41.54%
Computed Tomography	CT	158	22.01%	140	23.93%
Nuclear Scintigraphy	Scinti	14	1.95%	12	2.05%
Total		**718**		**585**	

First off all, for each information we search we must define dictionaries with all (as much as possible) of the terms that are representing that information (in our case, in French language). Thus, six modality dictionaries are created, based on the French version of MeSH (Medical Subject Headings - http://www.nlm.nih.gov/mesh/, http://www.chu-rouen.fr/ssf/arborescences.html). Each of the dictionaries is containing MeSH terms, inflected (plural) MeSH terms, synonyms of MeSH terms from the CISMeF terminology, inflected synonyms of MeSH terms and abbreviations representing each of the modalities. An extract from the **ultrasonography** modality dictionary (in French language), as well as the number of terms extracted for each of our six modality dictionaries are presented in Table 2.

Table 2

Modality Dictionaries

…
echographies,echographie.N+MeSH+TR:fp
ECHO,echographie.N+MeSH+TR
ultrasonographie,echographie.N+MeSH+TR+:fs
ultrasonographies,echographie.N+MeSH+TR+:fp
…

Angio	US	MRI	RX	CT	Scinti	**Total**
71	93	15	202	30	29	**440**

Once the dictionary created we use the linguistic INTEX/NOOJ environment [14] to extract all the modality related terms for the texts related to each of the 718 "image-text" pairs. Of course, actual results are obtained only for the 585 cases where image textual information exists in the source document. This approach is easily extensible to extract additional information by creating the corresponding dictionaries. Therefore, the same architecture is to be applied for the detection of ‹2›, ‹3›, ‹4› and ‹5›.

The text-region interpretation procedure we used for these experiments is modelled after the automatic indexing procedure proposed by CISMeF, for the indexing of entire medical documents, presented in detail in [3].

In the next section we are presenting the results obtained after applying these dictionaries for modality detection.

6. Results

For each text region, depending on the number of modality-related terms found, we can obtain either no modality decision, either one or several (not necessarily representing the same modality) decisions. A Majority Vote (MV) decision is then used to resolve some of the contradictions.

In Table 3 we present the results we obtain in the form of a confusion matrix. The "no MV text" column is representing the cases were no Majority Vote decision can be made and the "no text" column summarizes the images with no corresponding text. The *Retrieved* line is representing the number of decisions for each of the categories.

Table 3

Results

	Images with text							Images without text	Total
	Angio	US	MRI	RX	CT	Scinti	no MV decision		
Angio	**64**	0	0	0	0	0	1	39	104
US	0	**35**	0	0	0	0	2	3	40
MRI	0	0	**85**	0	1	0	2	0	88
RX	0	1	0	**233**	2	0	7	71	314
CT	2	0	1	0	**129**	0	8	18	158
Scinti	0	0	0	0	0	**11**	1	2	14
Retrieved	66	36	86	233	132	11	21	133	718
	564								

We can observe that there are 21 images where no MV decision was possible, 133 cases where there is no available text and only 7 misclassified images. A more detailed analysis for each class is presented in Table 4, by computing the mean Precisions, Recalls and F-measures for each of the modalities

(detailed information about the measures used in information retrieval can be found in [15]). We are especially interested in the F-measure rates because this measure combines the Precision and the Recall in a single efficiency measure (it is the harmonic mean of Precision and Recall).

Table 4

Precision/Recall/F-measure performances

	Precision (%)	Recall (%)	F-measure (%)
Angio	96.97	61.54	75.29
US	97.22	87.50	92.11
MRI	98.84	96.59	97.70
RX	100	74.20	85.19
CT	97.73	81.65	88.97
Scinti	100	78.57	88
mean	**98.46**	**80.01**	**87.88**

We obtain significantly high precision rates for all of the six modalities. This proves the existence of medical modality information in the document's image-related text regions, and that the proposed extraction and interpretation method is well suited for this problem. However, we note smaller rates for the recognition recall, due to incomplete image-text pairs in our extracted dataset. This is mainly caused by the absence of image-related textual annotations in the initial document or the absence of valid links between these annotations and the image. We observe the smallest recall rates for the Angio modality, for which more than 37% of the images have no corresponding extracted text. We consider as very important the fact that this approach, even though not capable of proposing a modality decision for each image (due to missing annotations), is still very accurate when image related text is available.

Combining this approach ‹c› with the other two implemented and tested in the MedIC architecture (‹a› and ‹b›) will most certainly improve the systems capability of accurately describing medical images.

7. Discussions and Conclusions

In this paper we evaluate whether the image-related text-regions of a medical document contain relevant medical information that can be extracted and used for automatic medical image description.

To prove the feasibility of this approach we first extract a set of 718 "image-texts" pairs from the resources indexed by CISMeF. These pairs are extracted from the most common document formats currently indexed by CISMeF. Giving the difficulties we encountered with the automatic extraction of these pairs, the PDF and HTML extractors are manually assisted/validated for the extraction of the 718 pair dataset.

We are focusing at evaluating the medical modality ‹1› significance of the text-regions. Dictionaries for each of the six modalities are created (using Fr-MeSH, CCAM thesaurus and CISMeF terminology), and the terms representing each of the six modalities are extracted using the INTEX/NOOJ platform. The good precision rates are indicating that this approach is viable if the text regions corresponding to the images are available and correctly extracted. A major advantage compared to other approaches of extracting image information (‹a› and ‹b›) is that this method is easily extensible to extract other information (‹1›→‹5›) by creating specific dictionaries. Furthermore, combining this decision with the other two approaches considered by MedIC (‹a› and ‹b›), we should be able to improve the global recognition recall rates, while maintaining high precisions. Further analysis of the significance of the extracted terms is in progress, by attributing different meanings to the terms located in image captions or paragraphs, or by weighting the terms with their distance to the image.

In the perspective of extending this architecture to seek other information than the modality, we must also note, that even though the architecture we presented in this paper is easily adaptable (by defining additional dictionaries), the variability of the information carried by on-line medical documents is to be considered (and better studied) before proceeding. Thus, we observed that the on-line published health documents (indexed by CISMeF) are covering a large scale of topics. The complexity level of the medical language, the terms and topics used are varying depending on the document's intended readers. The information carried by the image captions or the image-related paragraphs is equally varying. For instance, in didactic documents, where basic principles of medical imaging are usually introduced, the image-related text regions are mainly presenting the image modality, acquisition parameters, contrast agents used, details on the acquisition procedure or even introductory notions of physics for each acquisition principle (acoustics, X Ray, nuclear imaging, magnetic resonance). In contrast with these, for medical academic papers, diagnostic files (containing clinical cases), the information is more oriented on pathology and treatment, and thus, contain a lot less basic acquisition details. This document content variability is expected to influence de recognition performances of other medical information. Nevertheless, subjective evaluations of the extracted image-related textual annotations are showing the presence of an important number of terms related to anatomical regions, organs ‹2› and often pathology ‹4›. This is very significant because the pathology information is difficult to extract from the image it-self, the "recognition" of pathologies requiring a considerable amount of medical knowledge and complex reasoning mainly based on acquired experience.

As already stated, in the global architecture of MedIC module, the decision resulted from this approach ‹c› is intended to be combined with the decisions obtained using the information carried by the images ‹a› and ‹b›, to

obtain a more complete and accurate description of medical images. Used as a part of the automatic indexing of CISMeF resources, this module is designed to provide image semantic descriptions to the catalogue, to better assist users on their searches for quality heath-information on the Internet.

REFERENCES

[1] *T. Koch*, "Quality-controlled subject gateways: definitions, typologies, empirical overview. Online Information Review", **vol. 24**, no. 1, 2000, pp. 24-34

[2] *S.J. Darmoni, J.P Leroy, B. Thirion, F. Baudic, M. Douyère and J. Piot*, "CISMeF: a structured Health resource guide", in Methods Informatics in Medecine, **vol. 39**, no. 1, 2000, pp. 30-5

[3] *A. Névéol, A. Rogozan and S.J. Darmoni*, "Automatic indexing of online health resources for a French quality controlled gateway", in Information Processing and Management, **vol. 42**, no. 3, 2006, 695-709

[4] *F. Florea, H. Muller, A. Rogozan, A. Geissbuhler and S.J. Darmoni*, "Medical image categorization with MedIC and medGIFT", in Connecting Medical Informatics and Bio-Informatics - Medical Informatics Europe, 2006, pp. 3-11

[5] *F. Florea, A. Rogozan, A. Bensrhair, J.N. Dacher and S.J. Darmoni*, "Modality categorization by textual annotations interpretation in medical imaging", in Connecting Medical Informatics and Bio-Informatics - Medical Informatics Europe, 2005, pp. 1270–1275

[6] *W. Chu, C.C. Hsu, C. Cardenas and R.K. Taira*, "Knowledge-based image retrieval with spatial and temporal constructs", in IEEE Transactions on Knowledge and Data Engineering, **vol. 10**, no. 6, 1998, 872–888

[7] *E. El-Kwae, H. Xu and M.R. Kabuka*, "Content-based retrieval in picture archiving and communication systems", in Journal of Digital Imaging, **vol. 13**, no. 2, 2000, pp. 70–81

[8] *C.R. Shyu, C.E. Brodley, A.C. Kak, A. Kosaka, A.M. Aisen and L.S. Broderick*, "ASSERT: A physician-in-the-loop content based retrieval system for HRCT image databases", in Computer Vision and Image Understanding, **vol. 75**, no. ½, 1999, pp. 111–132

[9] *H.L. Tang, R. Hanka, H.H. Ip, K.K. Cheung and R. Lam*, "Semantic query processing and annotation generation for content-based retrieval of histological images", in International Symposium on Medical Imaging - SPIE Proceedings, **vol. 3976**, 2000

[10] *L. Long, S. Antani, D. Lee, D. Krainak and G. Thoma*, "Biomedical information from a national collection of spine x-rays: film to content-based retrieval", in Proceedings SPIE, **vol. 5033**, 2003

[11] *R. Marée, P. Geurts, J. Piater and L. Wehenkel*, "Biomedical image classification with random subwindows and decision trees", in Proceedings ICCV workshop on Computer Vision for Biomedical Image Applications, **vol. 3765**, 2005, 220–229

[12] *H. Muller, A. Rosset, J.P. Vallée and A. Geissbuhler*, "Integrating content-based visual access methods into a medical case database", in Studies in Health Technology and Informatics, **vol. 95**, 2003, pp. 480-5

[13] *T.M. Lehmann, M.O. Gűld, C. Thies, B. Fischer, M. Keysers, D. Kohnen, H. Schubert and B.B. Wein*, "Content-based image retrieval in medical applications for picture archiving

and communication systems", in Proceedings Medical Imaging, **vol. 5033**, 2003, pp. 440–451

[14] *M. Silberztein*, Dictionnaires électroniques et analyse automatique de textes: le système INTEX, Masson, Paris, 1993

[15] *C. VanRijsbergen, Information Retrieval, Ed. Butterworths, 2nd Edition, 1979*