Medical (Visual) Information Retrieval

Henning Müller 12

¹University of Applied Sciences Western Switzerland (HES–SO), ²University and University Hospitals of Geneva (HUG), Switzerland, henning.mueller@hevs.ch

Abstract. This text gives a broad overview of the domain of visual medical information retrieval and medical information analysis/search in general. The goal is to describe the specifics of medical information analysis and more specifically of medical visual information retrieval in this book of the PROMISE winter school. The text is meant to deliver an annotated bibliography of important papers and tendencies in the domain that can then guide the reader to find more detailed information on this quickly developing research domain. This text is by no systematic review in the field, so some citations might be subjective but should lead the reader to further publications. The given references will provide a solid starting point for exploring the domain of medical visual information retrieval.

Keywords: Medical information retrieval, content–based image retrieval, medical visual information retrieval

1 Introduction

Medical practice relies on data available on patients and usually tries to find evidence for or against specific diagnosis leading to further examinations or treatment [1, 2]. Decisions are thus often taken based on probabilities for or against a specific diagnosis. The more medical knowledge becomes available the more complex the relationships between the data and a potential outcome become. Modern medicine is thus increasingly producing data that can be treated by computers and the types of tests also change quickly over time. The amount of data produced per patient in modern hospitals has increased strongly over the past 30 years as has the amount of medical knowledge published in the scientific literature [3]. Medical imaging is in large part responsible for the data growth as modern tomographic devices produce ever thinner slices and also temporal sequences leading to an explosion of visual data produced. It is estimated that around 30% of world storage capacity is dedicated to medical imaging and that mammography in the United States alone accounted for over 2 Petabytes in 2009 [4] Analyzing such large amounts of data now requires computerized tools to remain efficient and particularly good processing infrastructures for computation [5]. Currently most use of the data is per patient but it has become clear that reusing the data to find connections and help solving cases with data of

other past cases can improve current care. Secondary use of medical data has thus been discussed many times [6, 7].

This article has the aim to introduce medical information retrieval in general as a domain but with a clear focus on retrieving visual information. Also for the retrieval of visual information many techniques from text retrieval are used and thus a general introduction is given and references to medical text retrieval. Then, the more detailed analysis is on the search for visual medical data.

Medical information retrieval has always been an active domain of information retrieval research [8] and many studies have been performed on the information searching behavior of physicians [9–11] showing that there are many information needs in clinical practice but that time is often too short for detailed search. Many physicians have regular information needs during clinical work, teaching preparation and research activities [9, 12]. Studies showed that the time for answering an information need with MedLine is around 30 minutes [11], while clinicians state to have approximately five minutes available [10]. Besides clear information needs there is also often a need to find similar cases, for differential diagnosis and also for cases–based reasoning [13].

Existing medical retrieval engines include the health on the net web page¹ for professional and also public access to health information. Professional access to the literature is given with the PubMed^2 search system that offers many access possibilities to the scientific biomedical literature including manual annotation of the articles with MeSH terms oragnized by the American National Library of Medicine. Medical search engines targeting radiologists but relying on text for research are Goldminer³ [14] and Yottalook⁴

Section 2 describes basic tools used to analyze medical texts, Section 3 details the main visual search techniques and approaches and finally Section 4 discusses the text critically with its main findings and ideas for future directions.

2 Medical Information Analysis and Retrieval

As said in the introduction, textual medical information retrieval is a mature domain with many techniques and applications available [8]. This domain deals with the analysis of medical texts in general and very often with extracting information from medical texts for further analysis. Natural language processing has for many years resulted in extracting information from medical texts [3, 15, 16] and sometimes mapping this information onto medical ontologies [17] to increase the value of extracted information. In general, medical ontologies have been created for many years to allow for higher quality coding of diagnoses, acts, and events of clinical practice. MeSH (Medical Subject Headings) is a terminology used to annotate PubMed scientific articles, UMLS (Unified Medical Language System) is a Metathesaurus containing a large number of terminologies and in

 1 http://www.hon.ch/

² http://www.pubmed.gov/

³ http://goldminer.arrs.org/

⁴ http://www.yottalook.com/

radiology RadLex is a standard terminoly for use in radiology reports [18]. All these terminologies have links between the items and allow for exploiting semantic analysis. The LinkedLifeData⁵ combines these terminologies and several others in a semantic repository that can also be used for extracting information from medical texts. Medical interoperability for data exchange can then rely on the many existing standards [19] to make sure that all partners in a particular health system actually understand the same standards and units when comparing information that have been shared. The National Library of Medicine in the United States is one of the largest actors in health data analysis and retrieval and much research is performed here. This includes text search engines [20] as well as approaches for the retrieval of images [21].

Many parts of medical information retrieval actually use the same principals as general text retrieval [22, 23]. The main particularities are really linked to a detailed analysis of terminologies and the sometimes big differences of medical language in several countries. This means that non–standard abbreviations are frequently used and in some languages latin forms are used, sometimes in combination or to replace normal language forms.

3 Medical Visual Information Retrieval

Several review articles give a much more complete view of the domain and the current tendencies than this somewhat subjective annotated bibliography [24– 26]. Early articles mentioning content–based medical image retrieval are [27–29].

3.1 Techniques

The basis for most visual retrieval applications are components for describing the images, or visual features, indexing and storage methods that allow for fast data access also with large databases, distances measures to compare two images or cases and then user interface components that allow presenting results to the user and interacting to optimize the shown results based on feedback obtained from users. Figure 1 shows this basic system layout with its components. Several of these components can include machine learning approaches such as the visual features or the distance measures that strongly depend on the type of data, and there are also many pre–processing steps that can be used to normalize the images for better comparison.

Most of the techniques used in medical image retrieval are broadly similar to techniques employed in non–medical systems. Detailed description of non– medical content–based image retrieval systems can be found in [30, 31]. One of the differences is clearly that medical text processing is quite advanced and that medical images can not really be analyzed without having their textual context. General images can often be analyzed with respect to simple objects or what is in the image whereas the context is also require to better integrate what the image is about or even what a picture invokes in people [32].

 5 http://www.linkedlifedata.com/

3.2 Applications

Visual medical information retrieval has remained for a long time a purely academic domain with most systems not even tested in clinical routine. One notable example for a small user test in a real clinical setting is [33] that used image retrieval for radiologists in diagnosing interstitial lung diseases showing that particularly inexperienced users gain from getting additional images through visual information retrieval. There are a few retrieval systems that are rather made for browsing in broad databases but these systems offer generally only a low general quality [29]. On the other hand many systems were developed for specific applications such as interstitial lung diseases [33, 34], spine images [35] or the liver [36]. A detailed overview of applications and also interfaces of medical image retrieval is given in [37]. A typical screenshot of a medical content–based image retrieval system can be seen in 2, showing the retrieved images to a visual query in a grid layout sorted by similarity.

Another screenshot can be seen in Figure 3 showing the Goldminer radiology search system. Here the images are shown in connection with the articles in which they appear highlighting the need for context and also the fact that most often cases are search for in clinical settings and not really single images without this context.

3.3 Evaluation of Visual and textual medical information retrieval

Within the Cross Language Evaluation Forum (CLEF⁶) an image retrieval task started in 2003 and a medical image retrieval task was added for 2004 [38]. This contest evaluates the quality of textual and visual information retrieval systems for medical texts with a focus on images and on multilingual retrieval. Varying data sets have been used over the ten years of its existence, starting with teaching files [39], then radiology journal articles and finally articles of the medical open

 6 http://www.clef-campaign.org/

Fig. 2. A screenshot of a typical medical image retrieval system showing the image results in a matrix ordered by similarity score and the diagnosis of the cases, including a link towards the case and the image in full resolution.

Fig. 3. Screenshot of the Goldminer radiology search engine with a list of images found for a textual search term in connection with the article in which they appear; visual search is not possible in this system.

access literature based on PubMed Central. In 2012 over 300'000 images were made available for retrieval. Three tasks are offered:

- image modality classification, so classifying each journal figure into one of 38 images modalities ranging from radiology modalities to biological image types and compound figures;
- image–based retrieval, meaning that the search target are single images; a search need is expressed through text in several languages and a few example images;
- case–based retrieval, meaning that the search target is a journal article that can be considered relevant for differential diagnosis of a given case that includes an anamnesis and images but no diagnosis.

More on the ImageCLEF campaigns and their outcomes can be found in [40, 41]. In general, several lessons have been learned over the past years:

- text retrieval is in general more stable and performs better than visual retrieval;
- combinations of textual and visual retrieval are delicate and the exact fusion often determines the quality of the final results with multimodal runs often obtaining the best results;
- approaches based on various types of visual words most often outperform all other approaches for visual retrieval;
- modality information and other classification–based outcomes can be used well to improve the retrieval results.

3.4 Challenges and Next Steps

Although content–based image retrieval is now well over 20 years old and also medical visual information retrieval over 15 years, still many challenges remain or have been uncovered in the past years as system get closer to clinical routine and physicians request new ways to navigate in the increasingly large data sets. One real challenge are clearly extremely large data sets or big data as very few of the current approaches scale well [42]. Approaches such as Hadoop/MapReduce exist but then still need to be adapted to the Terabytes produced in hospitals. What makes things worse in medical imaging is the fact that the regions of interest potentially relevant for a specific disease are often extremely small. Search by region of interest is one of the most frequently requested functionality of radiologists [43]. Ways to find out more about potential regions of interest in images can be eye tracking as seen in Figure 4. The images show clearly that the regions actually observed in detail are small and for most imaging types and suspected diseases we can probably create probability maps on whether it can potentially be a region of interest or not. This could potentially also reduce the amount of data to be treated and also transmitted to ease the burden of big data.

Another area that has been touched in ImageCLEF but will require much further research is *case-based retrieval* that also includes images in addition to free text and structured data. Case descriptions or journal articles often include

Fig. 4. A screenshot of a test with eye tracking equipment, show that regions of interest or volumes of interest are often rather small.

incomplete data and images taken for similar cases might vary between hospitals. The interactions and dependencies between images and clinical data are equally important and much has still to be learned. First publications on case–based retrieval exist in [44–46]. Something that could well help in this respect is a good annotation or coding of images that would allow the use of semantics. In radiology the RadLex standard [18] is an important step into this direction, and this plus the automatic extraction of semantic features can help much for the future. This would also allow to check data consistency and also contradictions using the large body of knowledge of the LinkedLifeData⁷.

Currently the by far strongest increase in medical imaging are multidimensional tomographic series, including a variety of modalities from CT (Computed Tomography), to MRI (Magnetic Resonance Imaging) and PET (Photon Emission Tomography) and combinations of these such as PET/CT. The increasingly thin slices create more detail and make viewing more difficult but offer many new possibilities for real 3D information retrieval [45, 47, 48]. Such solid 3D texture analysis can help to highlight regions of interest for physicians in volumes and make viewing easier. In the case of 4D data even the simple viewing becomes hard and so there is a read added value in making the data accessible in an easier way [49] 4D data sets can be 3D series with a time component or in the case of dual energy CT that creates 10 volumes of a body region imaged with varying energy levels that can potentially be useful in clinical application but go well beyond human vision.

Figure 5 shows an example of a 3D texture classification system that can be used as diagnosis aid. The different tissue types of the lung are shown in colored regions and can then be visualized in 3D allowing to explore the model or via a

⁷ http://www.linkedlifedata.com/

Fig. 5. A web–based interface of a system for diagnosis aid on lung diseases; the texture is classified into different classes and then shown in various 3D views.

standard color overlay in a grey scale slice view. This can help physicians viewing images quicker. In general visualization is an extremely hot topic and important in radiology to be able to analyze images quickly. The $Osirix⁸$ viewer is another important example allowing various views on the data and many plugins or tools for specific applications.

4 Discussion and Conclusions

This text gives a broad overview of medical information retrieval with a clear focus on the retrieval of visual information. The idea is to present an annotated bibliography and help getting a quick access to current developments in this field. This text is not a systematic review and thus some of the citations may be arbitrary but they do present current developments in the field and several classical review articles as well. Medical information retrieval has been a busy field for over 40 years and for medical visual information retrieval this has also been the case for 15 years. The way in which medicine moves towards an increasingly information rich field will make it necessary to develop new tools to make stored knowledge accessible and usable by physicians. Before images and their metadata will be fully integrated into information retrieval and clinical decision support I still expect several research rich years.

⁸ http://www.osirix-viewer.com/

5 Achnowledgments

This work was funded by the European Union in the context of the KHRESMOI (257528) and PROMISE (258191) FP7 projects. The PROMISE winter school was also funded by the ELIAS (ESF) project making this document possible.

References

- 1. Simel, D., Drummond, R.: The rational clinical examination: evidence–based clinical diagnosis. McGraw–Hill (August 2008)
- 2. Bui, A.A.T., Taira, R.K., Dionision, J.D.N., Aberle, D.R., El-Saden, S., Kangarloo, H.: Evidence–based radiology. Academic Radiology 9(6) (2002) 662–669
- 3. Hunter, L., Cohen, K.B.: Biomedical language processing: What's beyond pubmed? Molecular Cell 21(5) (Mar 2006) 589–594
- 4. : Riding the wave: How europe can gain from the rising tide of scientific data. Submission to the European Comission, available online at http://cordis.europa. eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf (October 2010)
- 5. Zhang, C., Sterck, H.D., Aboulnaga, A., Djambazian, H., Sladek, R.: Case study of scientific data processing on a cloud using Hadoop. High Performance Computing Systems and Applications Lecture Notes in Computer Science 5976/2010 (2010) 400–415
- 6. Safran, C., Bloomrosen, M., Hammond, W.E., Labkoff, S., Markel-Fox, S., Tang, P.C., Detmer, D.E.: Toward a national framework for the secondary use of health data: An american medical informatics association white paper. MIM 14 (2007) $1 - Q$
- 7. Elger, B., Iavindrasana, J., Iacono, L.L., M¨uller, H., Roduit, N., Summers, P., Wright, J.: Strategies for health data exchange for secondary, cross–institutional clinical research. Computer Methods and Programs in Biomedicine 99(3) (September 2010) 230–251
- 8. Hersh, W.: Information Retrieval A health and Biomedical Perspective. Springer (2003 – second edition)
- 9. Hersh, W., Jensen, J., Müller, H., Gorman, P., Ruch, P.: A qualitative task analysis for developing an image retrieval test collection. In: ImageCLEF/MUSCLE workshop on image retrieval evaluation, Vienna, Austria (2005) 11–16
- 10. Hoogendam, A., de Vries F. Robbé, A.F.S.P., Overbeke, A.J.: Answers to questions posed during daily patient care are more likely to be answered by uptodate than pubmed. Journal of Medical Internet Research 10(4) (2008)
- 11. Hersh, W.R., Hickam, D.H.: How well do physicians use electronic information retrieval systems? Journal of the American Medical Association 280(15) (1998) 1347–1352
- 12. M¨uller, H., Despont-Gros, C., Hersh, W., Jensen, J., Lovis, C., Geissbuhler, A.: Health care professionals' image use and search behaviour. In: Proceedings of the Medical Informatics Europe Conference (MIE 2006). IOS Press, Studies in Health Technology and Informatics, Maastricht, The Netherlands (August 2006) 24–32
- 13. Glasgow, J., Jurisica, I.: Integration of case–based and image–based reasoning. In: AAAI Workshop on Case–Based Reasoning Integrations, Menlo Park, California, AAAI Press (1998) 67–74
- 14. Jr., C.E.K., Thao, C.: Goldminer: A radiology image search engine. American Journal of Roentgenology 188 (2008) 1475–1478
- 10 H. Müller
- 15. Ruch, P., Baud, R., Geissbuhler, A.: Using lexical disambiguation and named– entity recognition to improve spelling correction in the electronique patient record. AIM 29 (2003) 169–184
- 16. Franz, P., Zaiss, A., nd Hahn Udo Schulz, S., Klar, R.: Automated coding of diagnoses – three methods compared. In: Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA), Los Angeles, CA, USA (November 2000)
- 17. Gobeill, J., Theodoro, D., Patsche, E., Ruch, P.: Taking benefit of query and document expansion using MeSH descriptors in medical ImageCLEF 2009. In: Working Notes of the 2009 CLEF Workshop, Corfu, Greece (September 2009)
- 18. Lanlotz, C.P.: Radlex: A new method for indexing online educational materials. Radiographics 26 (2006) 1595–1597
- 19. Müller, H., Schumacher, M., Godel, D., Khaled, O.A., Mooser, F., Ding, S.: Medicoordination: A practical approach to interoperability in the swiss health system. In: The Medical Informatics Europe Conference (MIE 2009), Sarajevo, Bosnia– Herzegovina (August 2009) 210–214
- 20. Ide, N.C., Loane, R.F., Demner-Fushman, D.: Application of information technology: Essie: A concept–based search engine for structured biomedical text. Journal of the American Medical Informatics Association 14(3) (2007) 253–263
- 21. Demner-Fushman, D., Antani, S., Siadat, M.R., Soltanian-Zadeh, H., Fotouhi, F., Elisevich, K.: Automatically finding images for clinical decision support. In: Proceedings of the Seventh IEEE International Conference on Data Mining Workshops. ICDMW '07, Washington, DC, USA, IEEE Computer Society (2007) 139–144
- 22. Salton, G., Buckley, C.: Term weighting approaches in automatic text retrieval. IPM 24(5) (1988) 513–523
- 23. van Rijsbergen, C.J.: Information Retrieval. Prentice Hall, Englewood Cliffs, New Jersey, USA (1979)
- 24. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medicine–clinical benefits and future directions. International Journal of Medical Informatics 73(1) (2004) 1–23
- 25. Akgül, C., Rubin, D., Napel, S., Beaulieu, C., Greenspan, H., Acar, B.: Content– based image retrieval in radiology: Current status and future directions. Journal of Digital Imaging 24(2) (2011) 208–222
- 26. Tang, L.H.Y., Hanka, R., Ip, H.H.S.: A review of intelligent content–based indexing and browsing of medical images. HIJ 5 (1999) 40–49
- 27. Lowe, H.J., Antipov, I., Hersh, W., Smith, C.A.: Towards knowledge–based retrieval of medical images. The role of semantic indexing, image content representation and knowledge–based retrieval. In: Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA), Nashville, TN, USA (October 1998) 882–886
- 28. Tagare, H.D., Jaffe, C., Duncan, J.: Medical image databases: A content–based retrieval approach. Journal of the American Medical Informatics Association 4(3) (1997) 184–198
- 29. Dahmen, J., Theiner, T., Keysers, D., Ney, H., Lehmann, T., Wein, B.: Classification of radiographs in the 'image retrieval in medical applications' — system (IRMA). In: 6th International RIAO Conference on Content-Based Multimedia Information Access, Paris, France (April 2000) 551–566
- 30. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content–based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12) (December 2000) 1349–1380
- 31. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys $40(2)$ (April 2008) 1–60
- 32. Jörgensen, C.: Retrieving the unretrievable in electronic imaging systems: emotions, themes and stories. In Rogowitz, B., Pappas, T.N., eds.: Human Vision and Electronic Imaging IV. Volume 3644 of SPIEProc., San Jose, California, USA (23–29 1999) (SPIE Photonics West Conference).
- 33. Aisen, A.M., Broderick, L.S., Winer-Muram, H., Brodley, C.E., Kak, A.C., Pavlopoulou, C., Dy, J., Shyu, C.R., Marchiori, A.: Automated storage and retrieval of thin–section CT images to assist diagnosis: System description and preliminary assessment. Radiology 228(1) (July 2003) 265–270
- 34. Depeursinge, A., Sage, D., Hidki, A., Platon, A., Poletti, P.A., Unser, M., Müller, H.: Lung tissue classification using Wavelet frames. In: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007., Lyon, France, IEEE Computer Society (2007) 6259–6262
- 35. Hsu, W., Antani, S., Long, L.R., Neve, L., Thoma, G.R.: Spirs: A web-based image retrieval system for large biomedical databases. International Journal of Medical Informatics 78(Supplement 1) (2009) S13–S24 MedInfo 2007.
- 36. Costa, M.J., Tsymbal, A., Hammon, M., Cavallaro, A., Sühling, M., Seifert, S., Comaniciu, D.: A discriminative distance learning–based CBIR framework for characterization of indeterminate liver lesions. In Greenspan, H., Müller, H., Syeda-Mahmood, T., eds.: Medical Content-based Retrieval for Clinical Decision Support. Volume 7075 of MCBR-CDS 2011., Lecture Notes in Computer Sciences (LNCS) (September 2011)
- 37. Depeursinge, A., Fischer, B., M¨uller, H., Deserno, T.M.: Prototypes for content– based image retrieval in clinical practice. The Open Medical Informatics Journal 5 (2011) 58–72
- 38. Clough, P., Müller, H., Sanderson, M.: The CLEF cross-language image retrieval track (ImageCLEF) 2004. In Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., eds.: Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Volume 3491 of Lecture Notes in Computer Science (LNCS)., Bath, UK, Springer (2005) 597–613
- 39. Hersh, W., Müller, H., Kalpathy-Cramer, J., Kim, E., Zhou, X.: The consolidated ImageCLEFmed medical image retrieval task test collection. Journal of Digital Imaging 22(6) (2009) 648–655
- 40. Müller, H., Clough, P., Deselaers, T., Caputo, B., eds.: ImageCLEF Experimental Evaluation in Visual Information Retrieval. Volume 32 of The Springer International Series On Information Retrieval. Springer, Berlin Heidelberg (2010)
- 41. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., García Seco de Herrera, A., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum). (September 2011)
- 42. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. In: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6. OSDI'04, Berkeley, CA, USA, USENIX Association (2004) 10–10
- 43. Markonis, D., Holzer, M., Dung, S., Vargas, A., Langs, G., Kriewel, S., Müller, H.: A survey on visual information search behavior and requirements of radiologists. Methods of Information in Medicine (2012) Forthcoming
- 44. Quellec, G., Lamard, M., Cazuguel, G., Roux, C., Cochener, B.: Case retrieval in medical databases by fusing heterogeneous information. IEEE Transactions on Medical Imaging 30(1) (jan. 2011) 108–118
- 12 H. Müller
- 45. Depeursinge, A., Vargas, A., Platon, A., Geissbuhler, A., Poletti, P.A., Müller, H.: 3D case–based retrieval for interstitial lung diseases. In: MCBR–CDS 2009: Medical Content–based Retrieval for Clinical Decision Support. Lecture Notes in Computer Science (LNCS), Springer (February 2010) 39–48
- 46. Zhou, X., Depeursinge, A., Stern, R., Lovis, C., Müller, H.: Case-based visual retrieval of fractures. In: International Journal of Computer Assisted Radiology and Surgery. Volume 5, Supplement 1., Geneva, Switzerland, Springer (June 2010) 11548/S162–S163
- 47. Depeursinge, A., Zrimec, T., Busayarat, S., M¨uller, H.: 3D lung image retrieval using localized features. In: Medical Imaging 2011: Computer–Aided Diagnosis. Volume 7963., SPIE (2011) 79632E
- 48. Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., Jacobs, D.: A search engine for 3d models. ACM Transactions on Graphics 22(1) (January 2003) 83–105
- 49. Johnson, T.R.C., Krauß, B., Sedlmair, M., Grasruck, M., Bruder, H., Morhard, D., Fink, C., Weckbach, S., Lenhard, M., Schmidt, B., Flohr, T., Reiser, M.F., Becker, C.R.: Material differentiation by dual energy CT: initial experience. European Radiology 17(6) (June 2007) 1510–1517