User-oriented evaluation of a medical image retrieval system for radiologists

ARTICLE *in* INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS · MAY 2015

Impact Factor: 2 · DOI: 10.1016/j.ijmedinf.2015.04.003

7 AUTHORS, INCLUDING:

Dimitrios Markonis

HES-SO Valais-Wallis

31 PUBLICATIONS **92** CITATIONS

SEE PROFILE

Markus Holzer

Medical University of Vienna

19 PUBLICATIONS **43** CITATIONS

SEE PROFILE

Célia Boyer

Health on the Net Foundation

79 PUBLICATIONS **570** CITATIONS

SEE PROFILE

Georg Langs

Medical University of Vienna

165 PUBLICATIONS **914** CITATIONS

SEE PROFILE

User-oriented evaluation of a medical image retrieval system for radiologists

Dimitrios Markonis[a], Markus Holzer[b], Frederic Baroz[c], Rafael Luis Ruiz De Castaneda[c], Célia Boyer[c], Georg Langs[b], Henning Müller[a]

[a]*University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland*

[b]*Medical University of Vienna, Vienna, Austria*

[c]*Health On the Net Foundation (HON), Geneva, Switzerland*

Keywords: Usability tests, User-centered design, Medical Informatics Applications, Content-based image retrieval.

Address for correspondence

Dimitrios Markonis, Msc,

HES-SO Valais

Rue du TechnoPole 3, 3960 Sierre, Switzerland

dimitrios.markonis@hevs.ch

tel +41 27 606 9033

Abstract

Purpose: This article reports the user-oriented evaluation of a text- and content-based medical image retrieval system. User tests with radiologists using a search system for images in the medical literature are presented. The goal of the tests is to assess the usability of the system, identify system and interface aspects that need improvement and useful additions. Another objective is to investigate the system's added value to radiology information retrieval. The study provides an insight into required specifications and potential shortcomings of medical image retrieval systems through a concrete methodology for conducting user tests.

Methods: User tests with a working image retrieval system of images from the biomedical literature were performed in an iterative manner, where each iteration had the participants perform radiology information seeking tasks and then refining the system as well as the user study design itself. During these tasks the interaction of the users with the system was monitored, usability aspects were measured, retrieval success rates recorded and feedback was collected through survey forms.

Results: In total, 16 radiologists participated in the user tests. The success rates in finding relevant information were on average 87% and 78% for image and case retrieval tasks respectively. The average time for a successful search was below 3 minutes in both cases. Users felt quickly comfortable with the novel techniques and tools (after 5 to 15 minutes), such as content-based image retrieval and relevance feedback. User satisfaction measures show a very positive attitude towards the system's functionalities while the user feedback helped identifying the system's weak points. The participants proposed several potentially useful new functionalities, such as filtering by imaging modality and search for articles using image examples.

Conclusion: The iterative character of the evaluation helped to obtain diverse and detailed feedback on all system aspects. Radiologists are quickly familiar with the functionalities but have several comments on desired functionalities. The analysis of the results can potentially assist system refinement for future medical information retrieval systems. Moreover, the methodology presented

as well as the discussion on the limitations and challenges of such studies can be useful for user-oriented medical image retrieval evaluation, as user-oriented evaluation of interactive system is still only rarely performed. Such interactive evaluations can be limited in effort if done iteratively and can give many insights for developing better systems.

Introduction

Images are an essential part in medical diagnosis and treatment planning. They are produced in quickly increasing quantities and also in with an increasing variety. Radiologists are often overloaded with the large amount of diagnostic images produced in hospitals that need to be read, and the ever-increasing number of image details (thin slices, temporal series, higher resolution) can put stress on the radiologist due to information overload. This creates risks to miss important structures or potential problems in the images. The medical literature often available on the Internet is also an important resource of visual medical information. However recent studies report that search for radiology images fails one out of four times [1]. The rapid growth of visual information available both in variety and quantity dictates the need for systems that facilitate quick access to relevant information. Medical information retrieval systems need to be able to handle real information scenarios in order to have an impact in radiology.

Much of the knowledge stored in images is little exploited at the moment because direct access to the visual image information (that is, the information that is contained in the visual content of the image represented by visual features) is rarely possible. Content-based image retrieval (CBIR) uses the visual content (such as shape, color and texture) of images or image regions as positive and negative examples to retrieve other images or cases that are related. Over the past 15 years, CBIR has been considered promising for assisting information search in the medical field and several systems have been developed [2] [3] [4] [5] [6]. However, most systems were rather technology-driven and very few applications have reached the end users for routine use or were integrated into the clinical workflow [7].

User-centered design (UCD) [8] has been used for several decades in industry [9] [10], but also in medical applications [11]. A few aspects of UCD have also been used for CBIR [12]. The concept behind this approach is to guide a system's design and development by investigating use case requirements and user feedback to improve the product's usability and the user experience. The key elements of UCD are described in the ISO (International Standardization Organization) standard for the human-centered design for interactive systems (ISO 9241-210, 2010) [13].

The first step in UCD of software applications includes investigation and understanding of the user requirements in order to identify the general design directions [14] [15]. User-centered evaluation is an im-

portant part of UCD in the early stages of the development [16] and needs to be seen as an iterative process throughout the development cycle [10] [11].

The assessment is often performed in the form of empirical usability tests in a number of target users to interact with the system in a lab environment or in a natural setting. Usability of the system is assessed with factors such as learnability, efficiency, effectiveness, memorability and satisfaction [16]. A survey on common usability testing techniques and tools is presented in [17]. The main methods for conducting such tests include thinking aloud execution of tasks, direct or recorded observation of the interaction, survey forms and log analysis. A more detailed description of aspects that need to be taken into account when designing a usability test is given in [18].

The number of users required for conducting user tests is another important aspect when designing a usability test. Early studies have reported that a single individual is not able to detect all usability problems but that 3-4 users are sufficient [19]. In [20] it is suggested that 5 users are enough, while other studies disagree, highlighting the need for larger user tests [21] [22]. The exact number of participants remains an open question, though in [23] it is proposed that 5 participants are indeed enough for each iteration of an iterative user-centered evaluation.

In this article a round of the user-centered evaluation of the Khresmoi[1] search engine is presented. This system aims at assisting general practitioners, the general public and radiologists in accessing trustable biomedical information. These three target groups have different search behavior, goals and information needs. Thus, the system is divided into subsystems, designed to correspond to the requirements of the target groups. Following the same concept, usability tests were designed and conducted separately, concentrating on domain-specific research questions.

This study focuses on the tests of the 2D image search prototype of the Khresmoi system that is designed to be used by radiologists. The system combines text and CBIR search for finding and navigating through scientific biomedical articles and the images they include. The prototype design is based on the investigation of the image use behavior of radiologists [1]. The backend of the system is based on the Parallel Dis-

[1] http://www.khresmoi.eu/

tributed Image Search Engine (ParaDISE) first used in [24] and the front end uses the ezDL interface [25].

The general research questions that the evaluation tries to answer are:

- Does the Khresmoi system improve current search for information in radiology (which is mainly patient–centered or using Google on the Internet for general information needs)?
- Does it cover finding answers to unmet information needs and to what extent are these covered?
- Which functionalities are more useful and which tools need to be improved, changed or added?

Materials and Methods

The system under evaluation was designed to respond to information needs of radiologists related to search for images and cases in the medical literature on the Internet. Therefore, an interface inspired by the state–of–the–art medical image and literature search engine interfaces [26] [27] served as a basis for the system prototype (see Figure 1).

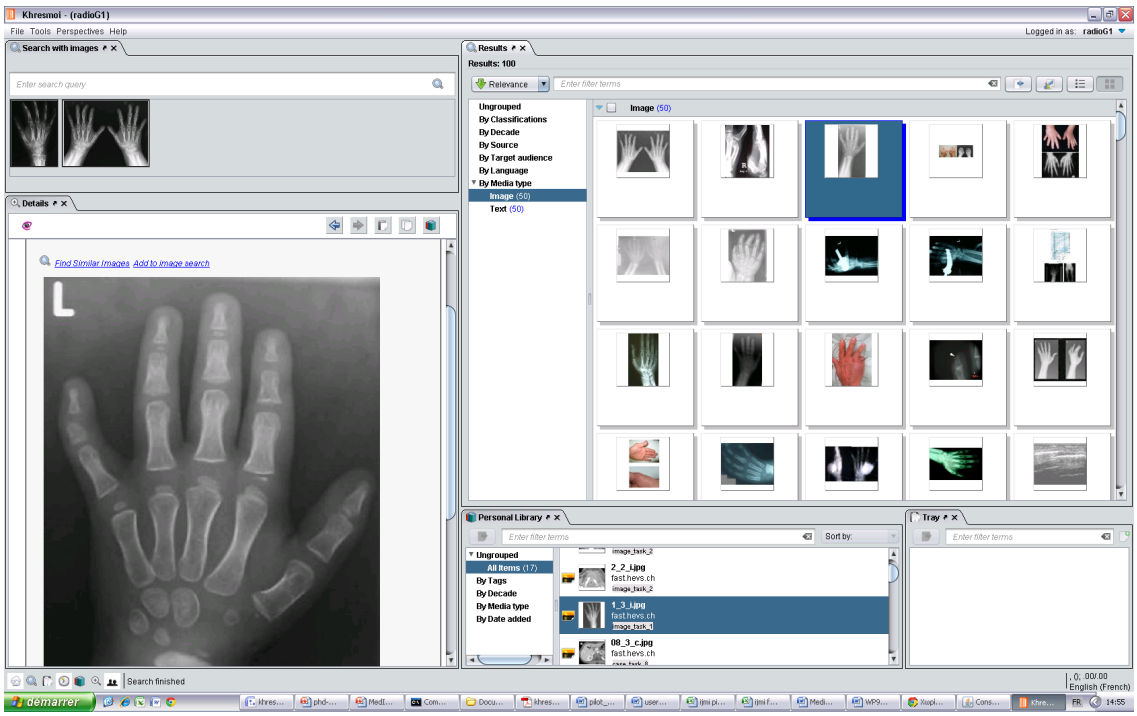

Figure 1 – Screenshot of the Khresmoi 2D image search prototype. The main tools are the query zone (top left), where user can use keywords and/or image examples as queries, the result list (top right) to quickly inspect results and the details view (bottom left) to view a selected result in more details. Additional tools are the personal library (bottom right), for storing interesting results and the tray for temporary storage.

The system allows query by text and/or image examples to retrieve images. Relevance feedback can be entered by the user by marking results as relevant or non-relevant to refine the search. Images can be selected to be displayed in full size and links to corresponding articles are provided. Interesting results can be stored in the personal library of the user for future use and investigation. Sharing of results and comments on documents among users is also supported.

The user-oriented evaluation process followed an iterative approach. Pilot user tests were performed to evaluate the basic aspects of the interface and the system functionalities [31]. This also helped detecting flaws of the user tests design and refining the study protocol. The study protocol was written down in a document and all persons involved in the tests (including the pilot test participants) discussed the text to make each test as independent as possible of the actual observer. In order to investigate the research questions described in the introduction, the following aspects were taken into account:

1. Success rate of information finding by radiologists using the Khresmoi system.
2. Time to find relevant information.
3. User satisfaction of the system performance.
4. Usability of the system.
5. Missing useful functionalities in the current version of the system.

The user satisfaction refers to the participant's opinion and emotions regarding the interaction with the system. Usability is connected to the ability to perform tasks in a straightforward way.

In this user study, the methods to acquire the above mentioned evaluation aspects needed to be decided upon. The final selection of methods, after being refined by the preliminary step of pilot user tests, is presented below:

- Participants were asked to perform information retrieval tasks for which at least one of the results is known. Therefore aspect number 1 could be evaluated.

- The time taken to fulfill each task was measured. This included formulating the query, inspecting the results and selecting the first relevant result. This method evaluated aspect number 2.
- Participants were asked to fill in a questionnaire about their experience when using the system. This allowed evaluating user satisfaction (aspect number 3) and detecting usability problems found by the participants. Questions were included that requested feedback and propositions for system improvement (aspect number 5).
- Participants were observed and video recorded while using the system (based on a signed informed consent that no one rejected). Possible system flaws or usability problems that were not consciously detected by participants were identified through this technique (aspect number 4).

Session outline

The user tests were conducted in the format of several one–to–one sessions, one participant performing the tasks and one observer present to facilitate the user test. The details of the session were also refined after the pilot tests by including and removing tasks, as well as modifying the time limitations. The final session outline is presented below:

1. Introduction to the Khresmoi project, the existing search system and the user test goals (5 minutes).
2. Tutorial video on the system tools and functionalities (5 minutes).
3. Demographic survey (5 minutes).
4. Introductory task, simple use of the tools (5 minutes).
5. Guided user tests in clear scenarios (30-40 minutes).
6. Survey on the satisfaction with the tools and functionalities (10 minutes).
7. Free possibility to use the system (5+ minutes).
8. Survey on the satisfaction with the system, free discussion (10 minutes).

The standardized introduction given by the test facilitator had as goal to help the participant understand the concept of the system and motivate to perform the test seriously. Then, the video demonstration of the system introduced the tools offered by the application. The introductory task was introduced after the pilot user tests because the video tutorial alone did not contain enough information for the user to get fa-

miliar with the tools available and the ways of using them. Throughout the session, the participant was being observed to identify potential shortcomings of the system. The observer was instructed to have a neutral attitude and was allowed to help only when the participant was blocked and could not proceed with a task. In order to limit the bias introduced by different observers, the test supervisor discussed with the test facilitators to define their role and their interaction with the participants precisely (e.g. in which cases they could assist the participant and what their observations should be focused on). Moreover, most of the tests were run in parallel sessions of three with the same person supervising the observers, to make sure that the protocol was strictly followed.

Task design, description and datasets:

During each session, the user was requested to perform several information seeking tasks. The design of the tasks took into account that they need to use most of the system tools and functionalities and cover the information needs of the target user group. They had to describe realistic scenarios that appear in clinical and academic workflows. Depending on the tasks different data sources were required. For this reason, the ImageCLEF2012 medical data set was used [28]. This data set contains more than 75,000 articles from PubMed Central open access journals and more than 300,000 images that are figures included in these articles. It represents a relatively realistic source for a medical literature search and especially for an initial test on the system's scalability and performance.

Two groups of information retrieval tasks were used: Three 2D image search tasks and two article search tasks. A subset of the ImageCLEF2012 medical image–based and case–based retrieval task topics was used respectively. The topics for the image–based task were selected after the log analysis of queries to a radiology image search engine [15]. Case-based topics consisted of cases included in an educational database [28] and each topic consisted of a description of the case and several images associated with the case. The guided scenarios of the user tests were based on these information retrieval tasks and included use of the various tools of the system, such as query–by–text, query–by–image–example, the personal library, the tray and others.

An example of the description of an image search task in its final form including query images is shown in Figure 2.

Task 3 (max 5 minutes):

1. Click on "Clear all content" from the File Menu.
2. Find as many images as you can that share the same diagnosis with the example images under the tag "image_task_3" in your Personal Library. You may mark results as relevant or non-relevant to relaunch the search.
3. Place the images you find into the Tray.

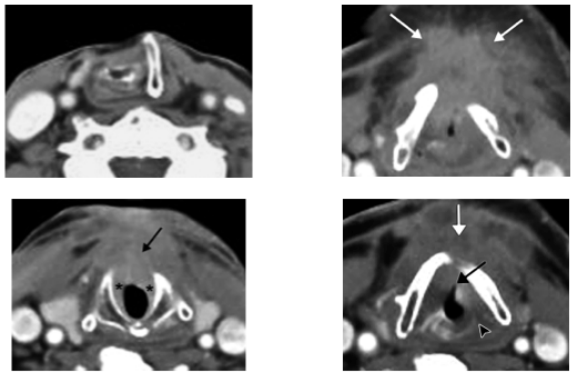

Figure 2 - Example of an image search task description and the query images given as examples.

An example description of the case-based search tasks and the images associated with the case are given in Figure 3:

Task 2 (max 7 minutes):

1. Click on the "Clear all content" from the File Menu.
2. You have the following case:

   "A 56-year-old woman with Hepatitis C, now with abdominal pain and jaundice. Abdominal MRI shows T1 and T2 hyperintense mass in the left lobe of the liver which is enhanced in the arterial phase."

3. You can find the images associated with the case placed in your Personal Library under the tag: "case retrieval_task_2".
4. Find as many relevant articles to the case above (that help in differential diagnosis) as you can and place them into the Tray.

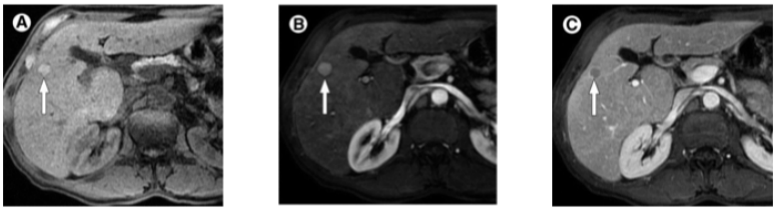

Figure 3 - Example of a case-based search task description and the query images given.

Session setup and tools used

The setup of the session included hardware and software preparation but also training sessions of the observer to get familiar with the recording tool and the study purpose. The hardware used in each session included two Windows-based computers one for the participant and one for the observer. The Khresmoi client was downloaded to the participant's computer and the recording software was installed on both computers. For observation and recording, the commercial software Morae[2] was used. The Morae software allows screen and face video recording of the participants and also remote online observing on a different computer. Upon start, Morae guides the user through the steps of the session, having all additional material, such as survey forms, task descriptions and instructions integrated and displayed on the

[2] https:// www.techsmith.com/morae.html

participant's computer screen. All of the survey answers, observer notes and recordings are saved in digital format, compatible with commonly used statistical packages for result analysis and presentation.

Three survey forms were used in this study. The initial demographics survey form was used to get information on medical experience and computer use of the participants. Finally, survey forms were used to evaluate the tools and the user satisfaction with the system. A combination of modified versions of the System Usability Scale (SUS) [29] and the Questionnaire for User Interaction Satisfaction (QUIS) [30] was used for the user satisfaction and usability survey forms. Open questions for providing comments on specific aspects of the system and suggestions for improvements were added. To get preliminary answers to the research goals, questions about the novelty, usefulness and intention of use of the tools were included.

At the end of each session the files containing the recordings, the answers to the surveys and the observer's notes were acquired and analyzed. The details of preparing, setting up and running a session were added into the study protocol document to ensure the reproducibility of the user tests.

Results

Two rounds of user tests were run in this study at the University Hospitals of Geneva and the Medical University of Vienna. The first set was a pilot user study to identify main system bugs and inconsistencies of the study protocol [31]. A next round of full user tests followed with a larger number of participants. The participants in the full tests are radiologists that volunteered to test the system. There was no overlap between participants in the pilot study and the full user tests.

Eleven persons (3 females, 8 males) participated in the full user test round. This number does not include the participants in the pilot user study and the interviews. They were all below 40 years old, with seven of them being below 30 and three between 31 and 35. Four persons were interns (participants 6, 7, 9 and 10), four were residents (participants 3, 5, 8 and 10), one associate professor in radiology (participant 1), one attending radiologist (participant 2) and one medical doctor with no specific radiology background (participant 11). Among the radiology specializations (participants could choose more than one field) the most common was thorax (3), radio oncology (3) and bone (2) while other chosen fields were echocardi-

ography, neuroradiology, cardiac, pediatric, general and emergency radiology. All of the participants reported frequent computer use (more than once a day) and search for medical information (7 reported more than once a day, 3 once a day and 1 once a week).

In two cases the participants did not perform all the tasks or answered all the questions due to technical difficulties, related to Internet being temporarily blocked in the hospital and recording software crashes and restarting. This resulted in 31 performed image search tasks out of 33 (11 participants × 3 tasks) and 19 article search tasks out of 22 (11 participants × 2 tasks).

The success rate was 87% (27/31) for image search tasks and 79% (15/19) for article search tasks. A task was considered successful if the user found at least one relevant result in the given time limit. A task was considered unsuccessful if the user could not find a relevant result in the time available (5 minutes for image search tasks and 7 minutes for case-based retrieval tasks).

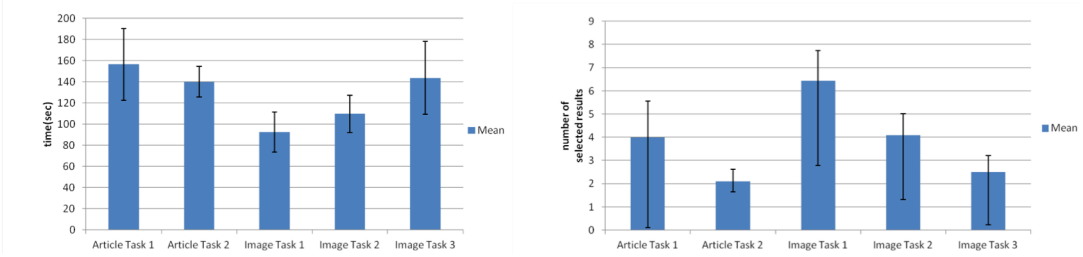

Figure 5 – Mean times and standard errors of the means of task completion (left) and number of selected results (right) per task.

The average time for finding the first relevant result during the image search tasks was 106 seconds. Again, this time included choosing image examples, investigating the results and judging a result as relevant. It includes only the cases when a relevant result was found. For case–based retrieval tasks the respective mean time was 150 seconds.

The mean number of results selected as relevant was 4 for the image search tasks and 3.1 for the case–based search. These numbers also include the cases for which no relevant results were found by the user. The mean and standard error of the mean of the retrieval times and the selected results number for each task are presented in Figure 5.

User satisfaction over specific system aspects and on the general use experience was measured using the answers on the survey questions on a Likert scale (where 1 was the lowest opinion and 5 was the highest), as can be seen in Figure 6. The full questionnaires are available in Appendix A. Figure 7 presents the mode for all the grades given by each participants. It is given as a qualitative measure of the general satisfaction of each participant as it removes extreme scores. For example, a mode of 4 or 5 indicates that the participant was satisfied with the majority of the aspects of the system usability. No common demographic variable could be found on participants ranking the system low (participants 2, 6 and 10).

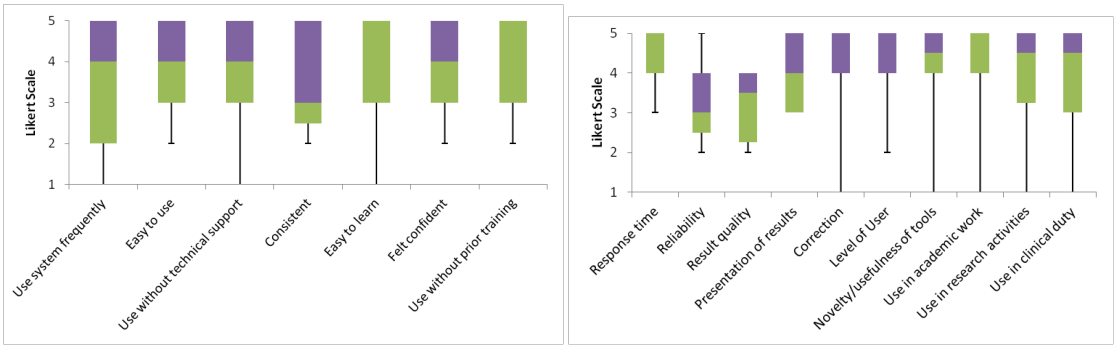

Figure 6 – Medians and grade ranges of user satisfaction on a Likert scale over specific aspects of the system (left) and general user experience (right). The line between purple and green boxes is the median value. The purple boxes show the standard deviation of the grades higher than the median and the green ones show the standard deviation of the grades lower than the median.

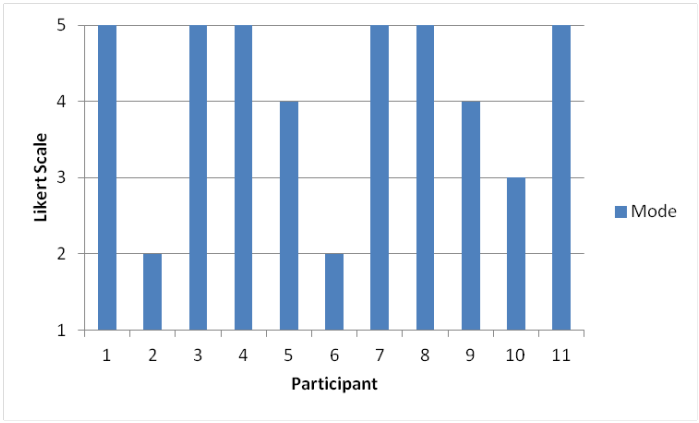

Figure 7 - Modes of grades given by each participant over the global satisfaction questions on a Likert scale.

The comments on the open questions were manually classified into frontend- and backend–related. Redundant comments were removed and the remaining comments were handed over to the development team so changes could be made in the system based on the comments. The most frequent comments can be summarized in the following points:

- Frontend
    - regarding querying, such as advanced text querying and relevant/non-relevant marking of images, the available options should be more explicit and easy to use;
    - basic and radiological–based image manipulation (such as changes in level window settings[3], zooming, etc.) of the selected and query images should be available;
    - results presentation and views; images should be presented in a grid as default and articles as lists.
- Backend
    - complaints were mentioned about CBIR returning many non relevant results; non relevant marking did not produce the desired results in modifying the query outcomes;
    - modality filtering was requested for more focused search;
    - propositions were made about finding articles using images only or using example articles as queries (”Find similar articles”).

The raw comments of the users are available as supplementary material in the online version of the paper as well as in [33].

Discussion

Lessons learned: pilot study

The pilot study was the first iteration of the user-centered evaluation, so focus was given on evaluating the user test design as well. It was found that the inclusion of a guided tutorial task after the video for the full user test protocol helped the users to feel comfortable with the system faster. The guided tutorial

[3] It is common in Radiology image viewer software to support using preset image values (e.g. brightness, contrast) for certain anatomical areas.

asked the user to perform very simple tasks using the tools in a first phase ad then corrected potential mistakes or difficulties. The use of a commercial recording and observation software such as Morae provides results in a unified digital format that is easy to transfer to statistical packages, to analyze and present in a meaningful way. On the other hand, the use of such a tool increases the hardware and software requirements and is prone to software crashes. Regarding the system, the users expected the system to give results that would correspond more to the keywords they entered in the query than to the image example. More detailed information about the pilot study's results can be found in [31].

Full user tests

The overall success in finding relevant images (87%) using the Khresmoi 2D image search prototype indicates an improvement over the percentage (~75%) reported in [1] on the image finding success rate of radiologists using existing tools. It should be noted that the results reported in [1] are based on self-assessment of the radiologists answering a survey and that users tend to overestimate their Internet use capabilities [32]. Case–based retrieval was shown to be a more challenging task (79% success rate), which was expected taking account empirical evaluation results [24].

The average time in the successful tasks for the participants to find a relevant result was less than 3 minutes for both types of tasks (1 minute 46 seconds for image retrieval and 2 minutes 30 seconds for article retrieval). This is also below the average estimated time reported in [1] (between 5 and 10 minutes) and indicates an added value in terms of time efficiency when using the Khresmoi system. Retrieval times can strongly vary as depicted in Figure 5, depending on the task. Other reasons may be the users' medical and computer experience. For the case-based retrieval tasks participants were not directly instructed to use images as queries. Some participants used the associated images only to investigate the case and not as query examples, querying using keywords as they are currently used to.

Results show that users were in general satisfied with the response time of the system (Figure 6). For text queries this means a response time of less than a second and for queries using image examples the time could be slightly above two seconds depending on the query. The reliability of the system was given a neutral median (3 with a range from 2 to 5). This can be due to the fact that the prototype while relatively stable had still a few minor bugs and inconsistencies also in the similarity ranking.

The quality of the results showed signs of improvement compared to the pilot study. A reason for this can be the modifications on the mixed text and visual retrieval ranking done before the full user test round. Another factor that must be taken into account is that the participants in the full user tests were in general more experienced with radiological analysis and used more advanced text queries. This resulted in more relevant documents and images returned by the system. However, results quality was still the weak point of the system and potential causes were identified. First, the CBIR performance was again mentioned to be returning many irrelevant results and users requested retrieved images to be of the same imaging modality as the query example. Another reason seems to be that the data set of biomedical images used contained only a small percentage of radiology images, limiting the retrieval performance of the radiology-related tasks.

In terms of user satisfaction with the presentation of the results, the feedback provided by the pilot study seems to have worked in a positive way. As requested, additional text information was included in the result list view and query terms were highlighted in the text. The users seemed to find the system novel and useful in practice giving a positive to strongly positive grade to this aspect (median of 4.5, with only one participant giving a grade below neutral). This was particularly encouraging, considering that the system is still in development and usefulness can be hidden by usability dissatisfaction. The activity that participants see as a likely area of use was academic work, which is along the design purposes. Eight radiologists out of eleven gave a mode above neutral over the global satisfaction aspects, as presented in Figure 7. This shows that the majority of the participants retained a positive or strongly positive attitude towards the proposed tools.

Much feedback was given by the participants in the open questions, post–test discussions and spoken comments while performing the tests. Some participants confirmed the outputs of the pilot user study while many new comments and propositions were introduced. On the graphical user interface aspect, the main comments were related to the image use, either requesting basic image manipulation features (which was also identified in the pilot tests but was not yet implemented for the full tests) or were about the image inconsistencies (e.g. drag and drop not being available for all views, detail views not being available for query images, non relevant marking being non–intuitive). Advanced text querying seemed to not be

straightforward and several participants either used advanced queries or at least asked about the availability. These facts may indicate that a more comprehensive interface would be useful for radiologists.

Overall, the system's concepts were appreciated, such as the connection of articles and images and the indexing of trustful sources. An improvement over the results quality would result in a system with even more practical use. Moreover, even though most of the tasks were successfully performed, the quantity and quality of resources returned in several scenarios was considered insufficient. More information and raw data on the user centered evaluation of the full Khresmoi radiology system, which includes also 3D image search and radiology report information exists in [33].

Study limitations

Radiologists have a tight time schedule and are difficult to recruit. This resulted in having a relatively low number of users for a quantitative analysis. For this reason, absolute quantitative results need to be taken with caution and serve mostly as indicators and as relative comparison. The user tests were performed in a lab room of a hospital and not in a room with diagnostic activity and standard viewing stations. This makes it difficult to assess the impact of the system on the actual clinical workflow of the radiologists. Moreover, because of the early stage in the system development and the low number of participants, no testing was performed to compare the system with other current solutions.

Conclusion

User-oriented design and evaluation is necessary for developing applications that correspond to realistic information needs and can have an impact on medical practice. Moreover, an iterative approach can provide more diverse results on the system usability evaluation.

Results show that young radiologists quickly feel comfortable in using new search tools, such as CBIR, relevance feedback and querying using complex statements. CBIR, despite its shortcomings in describing high-level concepts with simple visual features, can complement and improve text-based retrieval. Together with relevance feedback, it can facilitate quick query reformulation. The option to be able to filter

by specific medical image modality is often desired by radiologists. Grid representation of results with relevant text information seems to be preferred to vertical lists.

Several other points can also be addressed easily. Filtering by modality or image type can already filter out all non-radiology images to avoid having results of non-relevant image types. Separation of compound figures could also assist in avoiding mixed images in the results. The lack of relevant images can also be due to the limited representation of radiology journals in the 300'000 images chosen in the tests. The full PubMed set of 1.7 million images of 700'000 articles can also help in this aspect, as a larger database will have more relevant images and articles for the sometimes quite specific aspects. Currently, the retrieval system concentrates on 2D image search but an integration of clinical viewing of 3D image volumes and then a search for related medical articles would be even more useful from a clinical perspective.

Future work

The results of this study are useful as additional specifications for medical image retrieval system design and assist in avoiding potential pitfalls. Insights into the methodology for conducting meaningful user-centered evaluations are also provided in this text, so it should facilitate the creation of similar studies. The feedback obtained will guide future development of the Khresmoi system and all the points mentioned above are in the process of being addressed. Another round of tests for user-centered evaluation of the final Khresmoi system for radiologists is planned. Such an iterative approach can help bringing research prototypes much closer to real usefulness in clinical routine. Finally, a user test in the clinical environment could be performed to also measure the impact of a good retrieval system on diagnosis quality.

Acknowledgments

This work was partly funded by the European Union in FP7 via the Khresmoi project (grant agreement 257528).

References

[1] Dimitrios Markonis, Markus Holzer, Sebastian Dungs, Alejandro Vargas, Georg Langs, Sascha Kriewel, and Henning Müller. A survey on visual information search behavior and requirements of radiologists. *Methods of Information in Medicine*, 51(6):539–548, 2012.

[2] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content–based image retrieval systems in medicine–clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, 2004.

[3] Payel Ghosh, Sameer Antani, Rodney L. Long, and R. George Thoma. Review of medical image retrieval systems and future decisions. In *24th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6. 2011.

[4] Thomas M. Lehmann, Marc Oliver Güld, Christian Thies, Benedikt Fischer, Klaus Spitzer, Daniel Keysers, Hermann Ney, Michael Kohnen, Henning Schubert, and Berthold B. Wein. Content–based image retrieval in medical applications. 43:354–361, 2004.

[5] Henning Müller, Paul Fabry, Christian Lovis, and Antoine Geissbuhler. medGIFT — retrieving medical image by their visual content. In *World Summit of the Information Society, Forum Science and Society*, Geneva, Switzerland, 2003.

[6] William Hsu, Sameer Antani, L. Rodney Long, Leif Neve, and George R. Thoma. Spirs: A web-based image retrieval system for large biomedical databases. *International Journal of Medical Informatics*, 78(Supplement 1):S13–S24, 2009. MedInfo 2007.

[7] Alex M. Aisen, Lynn S. Broderick, Helen Winer-Muram, Carla E. Brodley, Avinash C. Kak, Christina Pavlopoulou, Jennifer Dy, Chi-Ren Shyu, and Alan Marchiori. Automated storage and retrieval of thin–section CT images to assist diagnosis: System description and preliminary assessment. *Radiology*, 228(1):265–270, July 2003.

[8] K. Vredenburg, J.Y. Mao, P.W. Smith, and T. Carey. A survey of user-centered design practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, pages 471–478, 2002.

[9] M. Hertzum. A case study of laboratory, workshop, and field tests. In *A. Kobsa and C. Stephanidis (Eds.), User interfaces for all, Proceedings*, volume 228, pages 59–72, 1999.

[10] A. Kaikkonen, A. Kekalainen, M. Cankar, T. Kallio, and A. Kankainen. Usability testing of mobile applications: A comparison between laboratory and field testing. *Journal of Usability*, 1:4–17, 2005.

[11] A. De Vito Dabbs, B.A. Myers, K.R. Mc Curry, J. Dunbar-Jacob, R.P. Hawkins, and A. Begey. User-centered design and interactive health technologies for patients. *Computers, Informatics, Nursing*, 27, 2009.

[12] J.C. Faga. Usability testing of a large, multidisciplinary library database: basic search and visual search. *Information technology and libraries*, 27:140–150, 2005.

[13] ISO 9241-210: 2010, Ergonomics of human-system interaction – part 210: Human-centered design for interactive systems. http://www.iso.org/iso/catalogue_detail.htm?csnumber=52075, 2010. Online; accessed 10–July-2014.

[14] Henning Müller, Christelle Despont-Gros, William Hersh, Jeffery Jensen, Christian Lovis, and Antoine Geissbuhler. 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, pages 24–32, Maastricht, The Netherlands, aug 2006.

[15] Theodora Tsikrika, Henning Müller, and Charles E. Kahn Jr. Log analysis to understand medical professionals' image searching behaviour. In *Proceedings of the 24th European Medical Informatics Conference*, MIE'2012, 2012.

[16] A. Holzinger. Usability engineering methods for software developers. *Communications of the ACM*, 48:71–74, 2005.

[17] C.J. Bastien. Usability testing: a review of some methodological and technical aspects of the method. *International Journal of Medical Informatics*, 79:18–23, 2010.

[18] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3:1–224, 2009.

[19] J. Nielsen and R. Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people*, pages 249–256, 1990.

[20] J. Nielsen and J.K. Landauer. A mathematical model of the finding of usability problems. In *CHI '93 Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, pages 206–213, 1993.

[21] J. Spool and W. Schroeder. Testing web sites, five users is nowhere near enough. In *CHI'01 extended abstracts on Human factors in computing systems, ACM*, pages 285–286, 2001.

[22] A. Woolrych and G. Cockton. Why and when five test users aren't enough. In *Proceedings of IHM-HCI 2001 conference*, pages 105–108, 2001.

[23] J. Nielsen. Alertbox. http://www.nngroup.com/articles/how-many-test-users/, 2012. Online; accessed 10–July-2014.

[24] Alba García Seco de Herrera, Dimitrios Markonis, Ivan Eggel, and Henning Müller. The medGIFT group in ImageCLEFmed 2012. In *Working Notes of CLEF 2012*, 2012.

[25] T. Beckers, S. Dungs, N. Fuhr, M. Jordan, S. Kriewel, and V.T. Tran. An interactive search and evaluation system. *Open Source Information Retrieval*, 9, 2012.

[26] Dimitrios Markonis, Adrien Depeursinge, Ivan Eggel, Antonio Foncubierta-Rodrguez, and Henning Müller. Accessing the medical literature with content–based visual retrieval and text retrieval techniques. In *Proceedings of the Radiological Society of North America (RSNA)*, November 2011.

[27] Charles E. Kahn Jr. and Cheng Thao. Goldminer: A radiology image search engine. *American Journal of Roentgenology*, 188:1475–1478, 2008.

[28] Henning Müller, Alba García Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner Fushman, Sameer Antani, and Ivan Eggel. Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In *Working Notes of CLEF 2012 (Cross Language Evaluation Forum)*, September 2012.

[29] J. Brooke. A quick and dirty usability scale. *Usability evaluation in industry*, 189:194, 1996.

[30] J.P. Chin, V.A. Diehl, and K.L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems, ACM*, pages 213–218, 1988.

[31] Dimitrios Markonis, Frederic Baroz, Rafael Luis Ruiz de Castaneda, Celia Boyer, and Henning Müller. User tests for assessing a medical image retrieval system: A pilot study. In *MEDINFO 2013*, 2013.

[32] A. J. A. M. van Deursen and J. A. G. M. van Dijk. Internet skills performance tests: Are people ready for ehealth? *Journal of Medical Internet Research*, 13(2), 2011.

[33] Dimitrios Markonis, Markus Holzer, Frederic Baroz, Rafael Luis Ruiz De Castaneda, Georg Langs, Celia Boyer, and Henning Müller. Report on the results of the initial user test of the radiology search system. Deliverable d10.2 of the khresmoi project, 2013. Available from: http://khresmoi.eu/assets/Deliverables/WP10/KhresmoiD102.pdf

Appendix A

1. *Are you ?*

- ○ Male
- ○ Female

2. *How old are you ?*

- ○ < 20
- ○ 20-30
- ○ 30-40
- ○ 40-50
- ○ 50-60
- ○ >60

3. *What is your native language ?*

- ○ English
- ○ French
- ○ German
- ○ Spanish
- ○ Czech
- ○ Other

4. *If you have chosen « other » please specify :*

5. *Your skills in English are ? <Scale>*

○ Basic, can comprehend simple issues ○ ○ ○ ○ Native language

6. *What is the highest position you have had in a medical service other than radiology ?*

7. *What is the highest position you have had in a service of radiology ?*
If non-applicable, enter N/A.

8. *If you have a work experience in a radiology service, how long have you been working in radiology ?*

○ N/A ○ 0-3 y ○ 4-6 y ○ 6-10 y ○ >10 y

9. *If you have a work experience in a radiology service, what field in radiology are you specialized in ?*

- ☐ Bone
- ☐ Nuclear radiology
- ☐ Radio-oncology
- ☐ Emergency radiology
- ☐ Thorax
- ☐ Interventional radiology
- ☐ Echography
- ☐ Other

10. *If you have checked « other » please specify :*

11. *Do you use a computer in your day-to-day life ?*

Never Once a month Once a week Once a day More than once a day

12. Do you use a computer for job or education related tasks ?

Never Once a month Once a week Once a day More than once a day

13. Do you use Google search ? <Scale + free text>

Never Once a month Once a week Once a day More than once a day

14. If you use other search engines, please specify below

15. Do you search the Internet for health related information ?

Never Once a month Once a week Once a day More than once a day

16. If yes, please specify the websites you use below :

17. Do you use Google image search ? <Scale + free text>

Never Once a month Once a week Once a day More than once a day

18. If you use other image search engines, please specify below :

19. Do you use Facebook ?

Never Once a month Once a week Once a day More than once a day

20. If you use other social media network, please specify below :

USABILITY OF THE SOFTWARE

1. ***I would like to use this system frequently.***

○ Strongly disagree ○ ○ ○ ○ Strongly agree

2. ***I found the system unnecessarily complex.***

○ Strongly agree ○ ○ ○ ○ Strongly disagree

3. ***The system was easy to use.***

○ Strongly disagree ○ ○ ○ ○ Strongly agree

4. ***I would need the support of a technical person to be able to use this system.***

○ Strongly agree ○ ○ ○ ○ Strongly disagree

5. ***The various functions in this system were well integrated, that is, the program works in a harmonious way which is logical to me.***

○ Strongly agree ○ ○ ○ ○ Strongly disagree

6. ***There was too much inconsistency in this system, that is, the program react in a way that I was not expecting and surprised me.***

○ Strongly agree ○ ○ ○ ○ Strongly disagree

7. ***I would imagine that most radiologists would learn to use this system very quickly.***

○ Strongly agree ○ ○ ○ ○ Strongly disagree

8. ***I found the system very awkward to use.***

○ Strongly agree ○ ○ ○ ○ Strongly disagree

9. ***I felt very confident on what I was doing, using the system.***

○ Strongly agree ○ ○ ○ ○ Strongly disagree

10. ***I needed to learn a lot of things before I could get going with this system. That is, the program requires a lot of training before an adequate use.***

○ Strongly agree ○ ○ ○ ○ Strongly disagree

11. ***Are there any tools that need to be improved/changed? If yes, how would you like them to be changed so that they will be more useful to your searches?***

12. ***Are there any new functionalities/tools that would like this search system to have?***

Screen Presentation

13. Reading characters on the screen

○ difficult ○ ○ ○ ○ easy

14. Add free comments

15. Presentation of images (e.g. size, position, additional information provided)

○ poor ○ ○ ○ ○ excellent

16. Add free comments

17. Quality of translation

○ poor ○ ○ ○ ○ excellent

18. Add free comments

19. Performing task is straightforward

○ never ○ ○ ○ ○ always

20. Add free comments

System capabilities (for 2D/Article search and 3D search)

21. Does the system respond quickly to your requests? Are the results delivered quickly enough?

○ too slow ○ ○ ○ ○ fast enough

22. Add free comments

23. Do you find the system reliable? Does it react the way you expect it to?

○ unreliable ○ ○ ○ ○ reliable

24. Add free comments

25. Are the results satisfactory? Do they match the queries you formulated?

Unreliable ○ ○ ○ ○ Reliable

26. Add free comments

27. Are the results well presented?.

Dislike how results are presented ○ ○ ○ ○ *Like how results are presented*

28. Add free comments

29. How easy is it to correct your mistakes, that is, undo, redo tasks ?

Difficult ○ ○ ○ ○ *Easy*

30. Add free comments

31. I think the system is appropriately designed for all levels of user (e.g. containing both simple and more advanced features in tools, for beginners and advanced in radiology respectively).

Strongly disagree ○ ○ ○ ○ Strongly agree

32. Add free comments

33. I think the system provides some tools and features that can be helpful in my work/research that are not available in the current tools I use.

Strongly disagree ○ ○ ○ ○ Strongly agree

34. Add free comments

35. I would use the 2D image and article search for academic work (preparation of lectures etc.).

Strongly disagree ○ ○ ○ ○ Strongly agree

36. Add free comments

37. I would use the 2D image and article search for research activities.

○ Strongly disagree ○ ○ ○ ○ Strongly agree

38. Add free comments

39. I would use the 2D image and article search during clinical work.

○ Strongly disagree ○ ○ ○ ○ Strongly agree

40. Add free comments