Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on the results of the large-scale user test of the radiology search system

Deliverable number	*D10.4*
Dissemination level	*Public*
Delivery date	*due 30.8.2014*
Status	*Final*
Authors	*Markus Holzer, Dimitrios Markonis, Georg Langs, René Donner, Roger Schaer, Johannes Hofmanninger, Markus Krenn, Erich Birngruber, Henning Mueller*

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Executive Summary

The KHRESMOI Radiology prototype has been redesigned according to the feedback obtained in the first round of the user–centered evaluation. The outcomes of that study were used to modify the system to include the most frequently requested functionalities and improve the aspects that were less satisfactory. This deliverable describes the final round user–oriented evaluation of the updated KHRESMOI system prototype.

The study protocol, the setup details and the results are presented and analysed. In total, 26 persons participated in the user test in Austria, Germany, Greece and Switzerland. The participants performed guided information retrieval tasks and answered questionnaires on the interaction with the system. They were observed while performing the tasks and their comments were collected. Usability aspects, such as effectiveness, efficiency and user satisfaction were recorded using the Morae observation and recording software and the results were analysed.

The results indicate that the KHRESMOI system has tools that have the potential to improve the current visual information search success rate and were found to be novel and useful in practice. Success rates in information finding and user satisfaction scores were higher than the ones from year 2 user–centered evaluation. The modifications made after that evaluation have been found beneficial for the system's usability. Users proposed functionalities, such as preset window levels for the 3D prototype or auto–complete for semantic image search. Searching for similar articles was also suggested by some participants.

In conclusion, KHRESMOI Radiology could be a useful tool for the clinicians. There are improvements to be made for this tool to become a real work application, such as the modality classification accuracy or the 3D retrieval precision, especially for cases that contain multiple pathologies. Having a less crowded interface could also improve the user experience. In general, long term user tests in a clinical environment could potentially expose more about the usability of the system and find more about additional functionalities required.

Table of Contents

List of Figures

Abbreviations

CBIR	Content–based image retrieval
UCD	User–centered design
SUS	System Usability Scale
QUIS	Questionnaire for User Interaction Satisfaction
ROI	Region of Interest
PACS	Picture archiving and communication system
ID	Identifier

1 Introduction

User–centered design and development is a crucial part in the KHRESMOI project. The first round of the KHRESMOI Radiology prototype user tests [6] indicated that the ideas behind the system design were appreciated by young radiologists. Novel functionalities offered, such as the linking of images and articles, were found to be helpful, while others, such as content–based image retrieval (CBIR), were reported to need improvements. The trustworthiness and the quality of the external sources that were indexed by the system were satisfactory. However, at the same time there were indications that the amount of indexed data was insufficient.

The findings of that evaluation round were used to redesign and improve the system. The final prototypes included improvements such as the indexing of a larger external resources corpus, the inclusion of automatic image modality classification and the revision of the CBIR algorithms [4]. The 3D retrieval implementation achieved response times of less than 5 seconds and the 2D image search prototype included basic image manipulation similar to the 3D prototype. Integration of the two prototypes was obtained by extending 3D searches to external sources based on the consensus of the top results [5].

In this report, the final round of the user–centered evaluation of the KHRESMOI Radiology prototype is presented. The general research questions that the evaluation tries to answer are very similar to the ones of the first evaluation round:

- Does the KHRESMOI system improve current search for information in radiology (which is mainly patient–centered or using Google on the Internet)?
- Does it cover unmet information needs and to what extent?
- Are the functionalities that were added after the first evaluation round useful and which tools need to be improved, changed or added?

2 Methods and Materials

This section describes the methodology followed by designing, setting up and running the user tests. A slightly modified version of the study protocol of the first round of the user tests (described in [6]) was used in this study. It is given in detail again in this section for completeness reasons.

2.1 User study protocol

In order to investigate the research questions described in the introduction section, the following aspects were taken into account:

1. Success of information finding by radiologists using KHRESMOI.
2. Time to find relevant information using KHRESMOI.
3. User satisfaction of the KHRESMOI system performance.

4. Usability of the KHRESMOI system.
5. Open challenges for future research.

The methods of the above mentioned evaluation aspects were kept the same as in [6] to allow for possible comparisons:

- Participants were asked to perform information retrieval tasks for which at least one of the results is known. Therefore aspect no.1 could be evaluated.
- The time taken to fulfill each task was measured. For tasks whose time was fixed, the time taken to find the first relevant result was measured, instead. This method evaluated aspect no.2.
- Participants were asked to fill a questionnaire about their experience of using the system. This allowed evaluating user satisfaction (aspect no.3) and detect usability problems found by the participants. Questions were included that requested feedback and propositions for system improvement (aspect no.4).
- Participants were observed and video recorded while using the system. Possible system flaws or usability problems that were not consciously detected by participants were identified through this technique (aspect no.4).

The user tests were conducted in the format of one–to–one sessions, one participant performing the tasks and one observer 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. Demonstration of the system tools and functionalities by the observer (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 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 introduction by the observer was intended to help the participant understand the concept of the system and provide motivation to do the test. Then, the demonstration of the system introduced the tools offered by the application. The introductory task was introduced after the pilot user tests of the first evaluation round to give the user some time to get comfortable with

the interface commands. In order to decrease the run–time of a session the number of tasks was reduced from 10 (which was the first evaluation round task number) to 7. Throughout the session, the participant was being observed by the observer 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.

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 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.

At the end of each session the file containing the recordings, the answers to the surveys and the observer's notes were acquired. The details of preparing, setting up and running a session were added into a document to ensure that the experiment can be reproduced under the same conditions. This document of instructions can be found in Appendix D of [6].

2.2 Description of prototypes

In this section the 2D image/article search and the 3D image search prototype are described.

2.2.1 Description of 2D image/article search prototype

The description of this prototype is given in Deliverable D2.6 [5]. However, the semantic image search feature was added since that report and will be described here for completeness reasons. It is presented in Figure 1. The semantic image search query tab contains dropdown lists for modality, anatomy and pathology RadLex terms. Once one or more terms are selected in at least one field, a search can be initiated. The results are shown and relevance feedback can be used to refine the search. Moreover a list of related terms is presented in the Details View, that can be used to further restrict the search.

2.2.2 Description of 3D image search prototype

The user interface used for the 3D Image Search prototype consisted of two perspectives.

As starting point the index perspective shown in Figure 2 displays all the available cases in the database. The user is able to filter each column with respective values in order to search for specific datasets and it is also possible to filter the medical findings of the datasets. If a volume is selected by left clicking on it in the index view, its center slice is displayed on the right side together with the medical finding and marked ROIs (if any). Clicking on the slice initiates the loading of the volume which is then displayed on top of the current perspective (Figure 3), where the user can manipulate the volume by brightness, contrast, pan and zoom. The user is able to adept brightness and contrast left clicking on the volume, which enables the so called windowing function, where the user can change the brightness by moving the mouse forward/backward and the contrast by moving the mouse left/right. A second left click disables the windowing mode. By right clicking on a slice of the volume the ROI marking mode can be enabled. The volume can be closed by pressing on the close button and a search can be initiated by clicking on the search button.

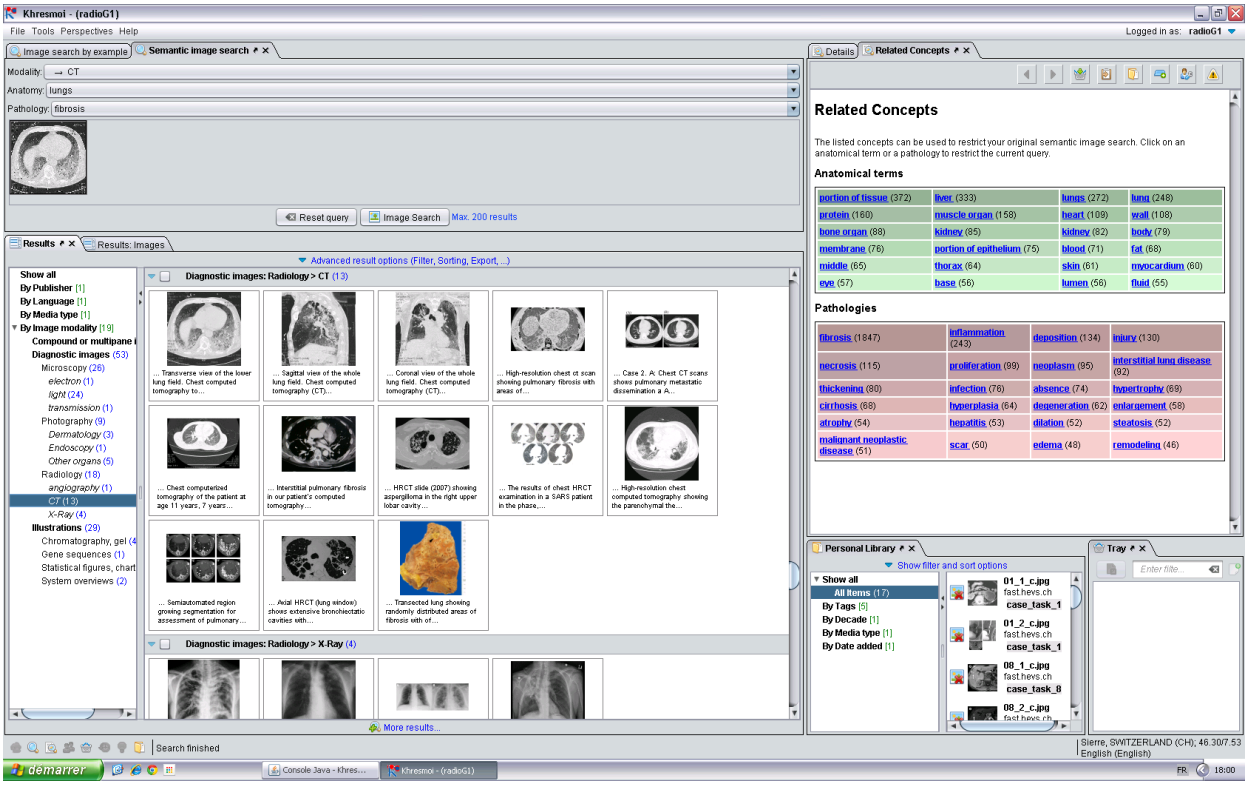

Figure 1: The semantic image search perspective.

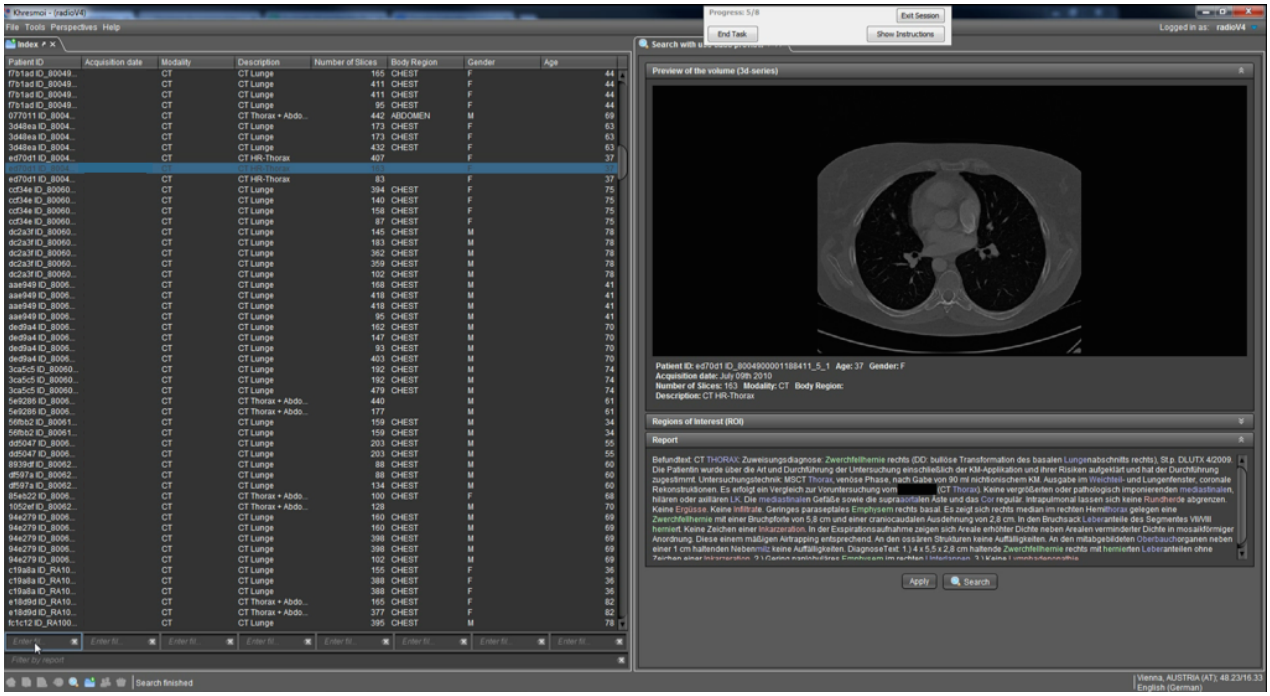

Figure 2: The index view of the 3D Image Search prototype.

Figure 3: The ROI marking (full-volume-view) overlaid on the index view of the 3D Image Search prototype.

The system now switches into the result perspective, which is divided into three windows as shown in Figure 4. On the left the query volume is displayed, while in the center the result thumbnails of the best matching slice and report summary are shown. In the center window three additional tabs are created containing the results of the integrated 2D prototype. These contain results matching a generated query string (visible in top left of the result perspective) in terms of article search, 2D image search and RadioWiki search. On the right side the details of the selected result are presented. If a volume is selected in the center window, its thumbnail and medical finding are displayed in this window. The user can now load the full volume and the generated overlays by clicking on request images.

Clicking on the perspective menu in the top left allows the user to manually switch between the different perspectives.

2.3 Tasks and data sets

As mentioned in Section 2.1, the user was requested to perform several information seeking tasks during a session. The design of the tasks took into account that they need to use most of the system tools and functionalities, particularly those modified/added (after the first evaluation round) ones. They had to describe realistic scenarios that appear in clinical and academic workflows. Depending on the tasks and the used subsystem, different data sources were required. The tasks and data sets used are described in sections 2.3.1 and 2.3.2.

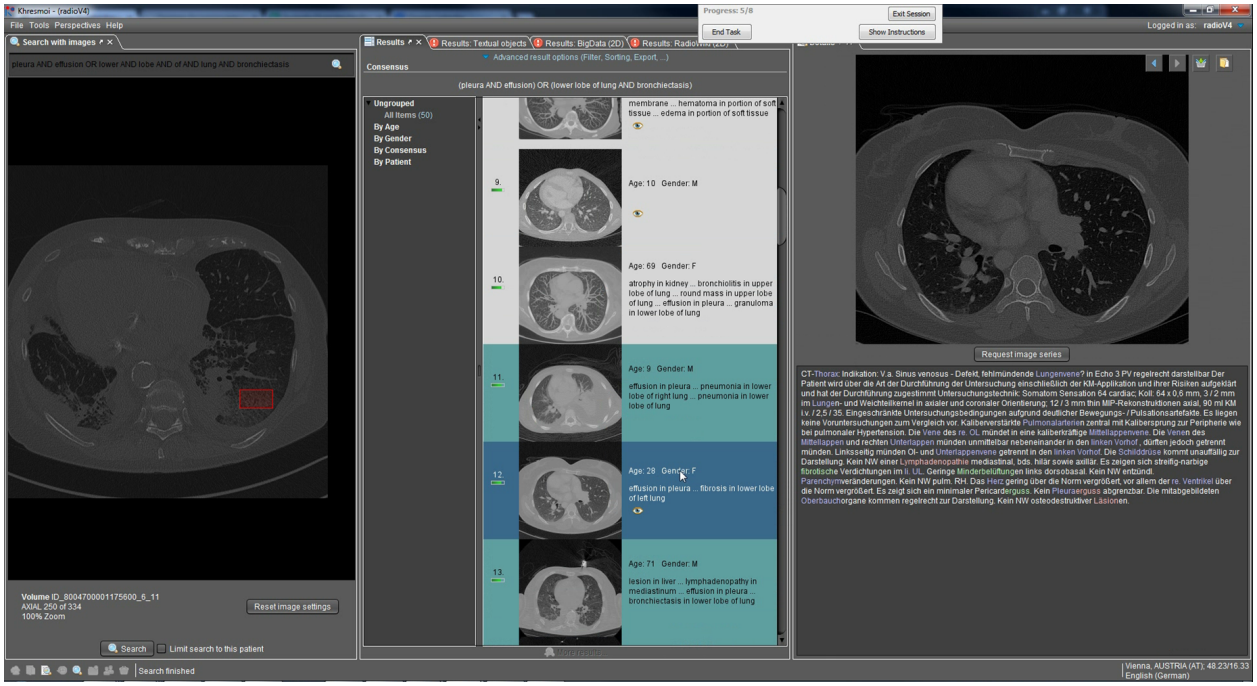

Figure 4: The result view of the 3D Image Search prototype.

2.3.1 2D image/article search

From the feedback of the first round of evaluation it became apparent that the ImageCLEF2012[1] [8] indexed data amounting to 75,000 articles and 300,000 images was not fully adequate for the information seeking needs of radiologists. For this reason the PubMed Central dataset, consisting of 500,000 articles and more than 1,700,000 images, was used for the evaluation of the KHRESMOI system in terms of radiology–related information search as external sources.

Two groups of information retrieval tasks were used: Two 2D image search tasks and one article search task. For the first image search task and the article search task, an ImageCLEF2012 medical image–based and a case–based retrieval topic were used respectively. In ImageCLEF the image–based topics were selected after the log analysis of queries to a radiology image search engine [9], while case-based topics consisted of cases included in an educational database [8].

The tasks of the user tests were based on these information retrieval tasks and included the use of the various tools of the system, such as query–by–text, query–by–image–example, the personal library, the tray and others. More importantly, use of the newly added features, such as filtering by modality and image manipulation was included in the instructions of tasks. The full task descriptions can be found in Appendix C of [6].

An extra image search task was created in order to evaluate an experimental feature, that included image search using the semantic web. The instructions of this task asked the participant to use the semantic image search perspective for searching for CT images of fibrosis in the lungs.

[1] http://www.imageclef.org/

2.3.2 3D image search

The focus of the final user tests of the 3D prototype is on the pathology retrieval. Therefore the anatomical dataset was discarded and the pathology dataset was increased from 117 to 1163 Lung CT volumes. Also the number of labeled pathologies was increased to atelectasis, bullae, emphysema and ground-glass.

In order to evaluate the quality of the prototype the users were given two similar tasks. For each of them a case was randomly chosen from six example cases. Both of the tasks included the following:

- The unique ID and the pathology specific slice of the volume.
- The assumption that there is no report available for this volume.
- If the user knows one or more of the pathologies visible in the given slice the goal is to try to verify it/them using the image search feature and also find a better visual representation of the pathology that could be shown as an example.
- If the user does not know what the visible pathology is, the image search feature should be used to find out what it might be and again the user should try to find a better visual representation of the pathology that could be used as an example.
- Results that were useful would be stored in the personal library.

The focus of the first task was to also take a specific look at the consensus (green background) and non-consensus (white background) suggestions of the algorithm that were indicated in the result view. This was done in order to identify if the automatic division is useful for the physicians.

In the second task the users were advised to specifically look at the additional 2D/Article search tabs as the task was also about finding out if the integration of the 2D prototype was useful for the physicians.

The six example cases were given by the unique ID of the volume and a specific slice and contained the following pathologies:

- Centrilobulary emphysema
- Bullae emphysema
- Fibrosis
- Bronchiecstasis
- Ground-glass
- Honeycombing

The variety and randomization of pathologies ensured that the user satisfaction is not based on specific pathology retrieval.

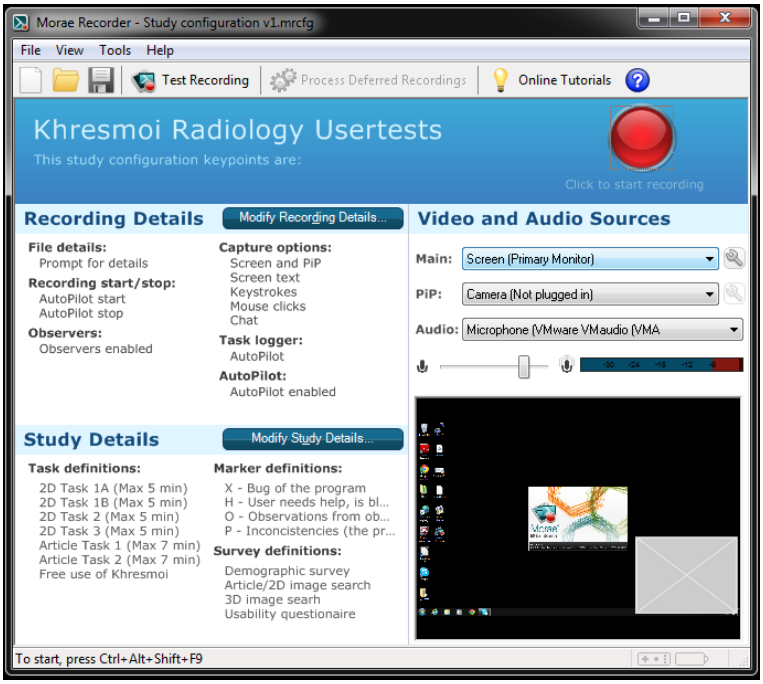

Figure 5: The interface of the Morae Recorder software, which was installed on the participant's computer. A common study configuration file is used for all the users and the recording starts by pressing red button.

2.4 Recording material

For observation and recording, the commercial software Morae mentioned in [1] was used. This software allows screen and face video recording of the participants (Figure 5) and also remote online observing on a different computer. Moreover it facilitates conducting the user tests by displaying the introduction and task description text as well as surveys on the participant's computer screen. All the survey's answers, observer's notes and recordings are saved in a digital format which is compatible with commonly used statistical packages for result analysis and presentation.

2.5 Survey forms

Four 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. Two survey forms were used to evaluate the subsystem's tools and functionalities usability and one to evaluate user satisfaction with the global system. A combination of modified versions of the System Usability Scale (SUS) [2] and the Questionnaire for User Interaction Satisfaction (QUIS) [3] was used for the user satisfaction and usability survey forms. Open questions for providing comments on specific aspects of the system, especially the newly added features 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. The survey forms can be found at the Appendices A and B of [6].

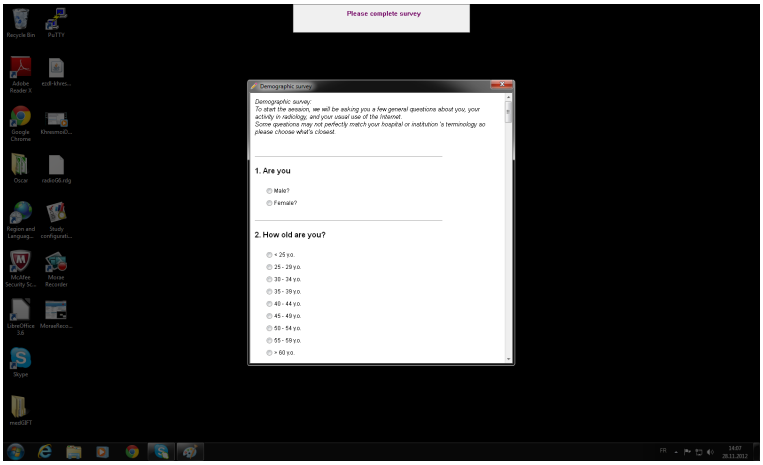

Figure 6: Screenshot of filling in a survey form using the Morae software.

3 Results

The final user tests took place at the University Hospitals of Geneva, Switzerland, the Medical University of Vienna, Austria, the University Hospital of Freiburg, Germany and the University Hospital of Larissa, Greece. In total 26 users participated in the tests, split into 17 male and 9 female. The age distribution shown in Figure 7 ranges from 20 to 60 years. There are 7 different native languages within the participants, with a majority of eighteen speaking German, while two speak Greek and one for each of the following languages: Hungarian, Romanian, Macedonian, Croatian, French. They self-assess their English skills to a median of 4 on a scale of basic-1 to native-5 as shown in Figure 8. Sixteen of the participants were residents, two where consultants, two associate professors and one for each of the following positions: general physician, junior consultant, vice chairman, head of unit and attending radiologist. The distribution of their years working in radiology is presented in Figure 9. Almost half of the participants specialized in the fields of thorax, emergency radiology or intervention radiology. The full distribution is shown in Figure 10. The demographic information about usage of computer and internet is summarized in Figure 11. All of the participants use a computer in their day-to-day life as well as for their job or education related tasks. Also search engines (Google) are used more than once a day. The internet is used to search for health related information at least once a day and by 75% more than once per day. Most of the participants (eighteen) use the Google image search at least once a day and eight of the participants do not use any social network.

The user satisfaction is analyzed over key general aspects of the system and is presented in Figure 12. Results are compared to those of the previous prototype evaluation where possible. The median for the question about intention to use the system frequently increased from 4 to 5. The same median was obtained for straightforward design, the easiness to learn how the prototype works and that there is no special training required. The consistency of the prototype increased from a median of 3 to 4. A median of 4 was also obtained for the remaining satisfaction criteria.

In order to assess the global satisfaction of each participant the mode over the general satisfaction question was measured. The frequency of the mode is an indicator for the consistency.

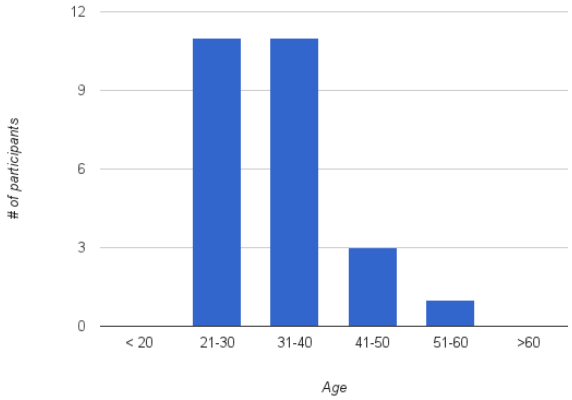

Figure 7: Age distribution of the user test participants.

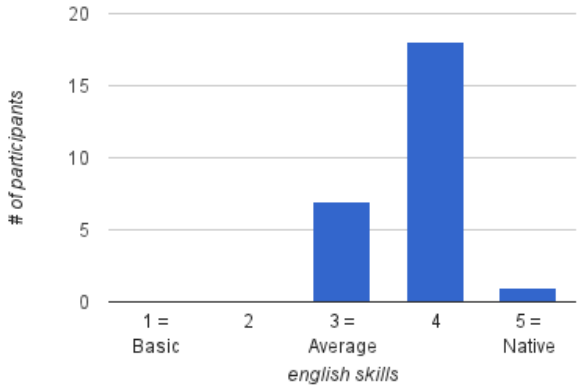

Figure 8: Self-assessment of English skills by participants.

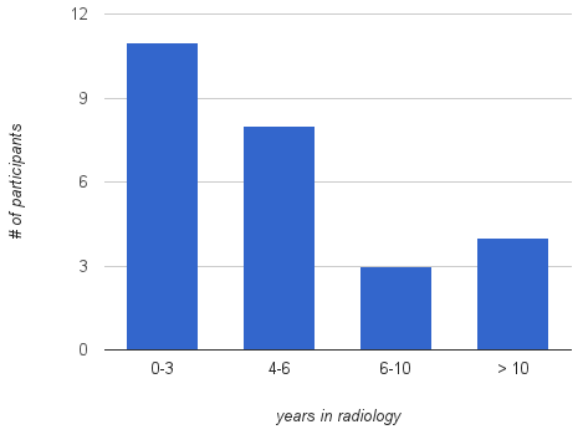

Figure 9: Time of service in the field radiology.

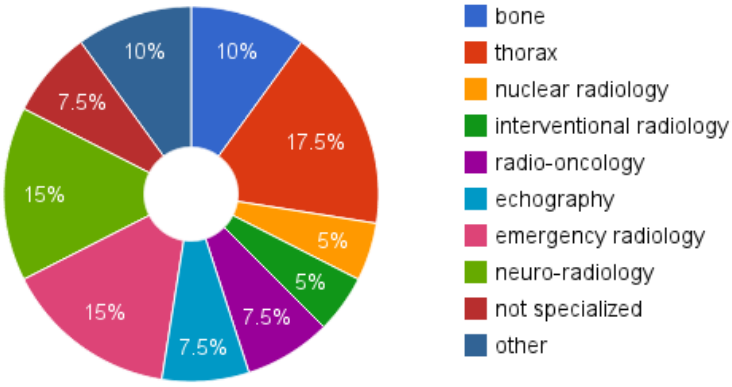

Figure 10: Field in radiology of the participants.

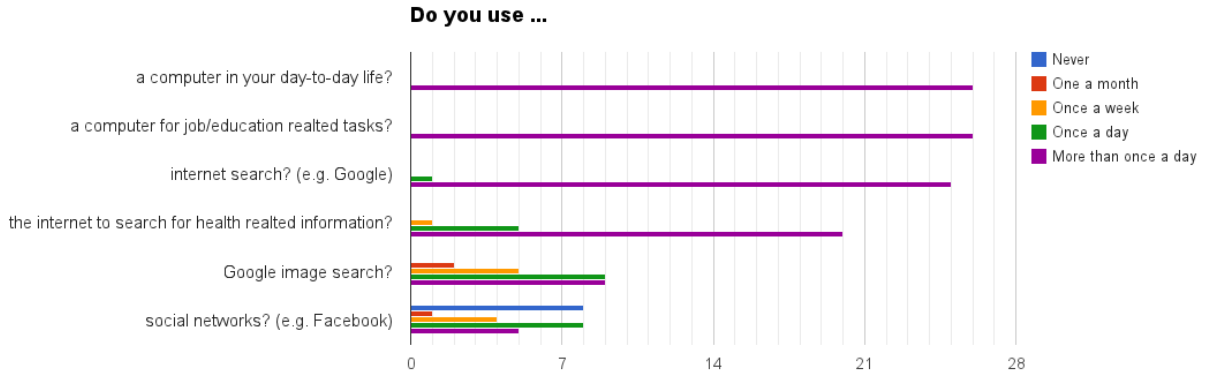

Figure 11: Demographic information about computer and internet usage of the participants.

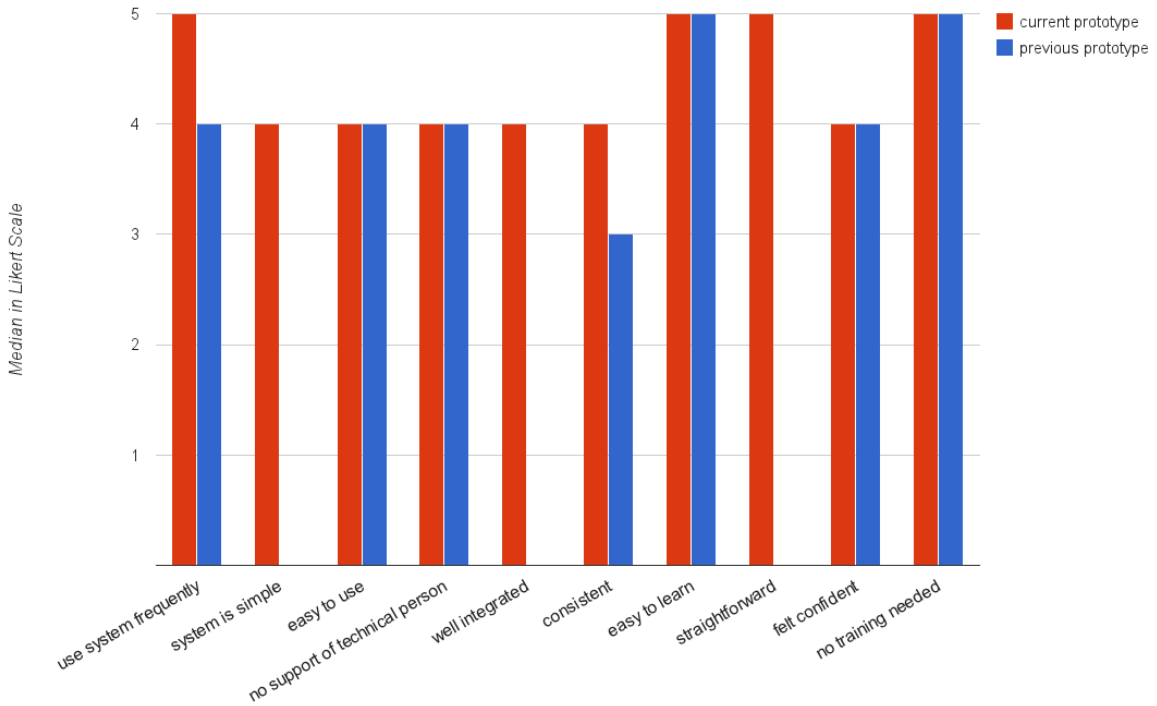

Figure 12: Median values of measuring general user satisfaction about the system in Likert scale (1=strongly negative, 5=strongly positive) compared to the median values of the first prototype evaluation.

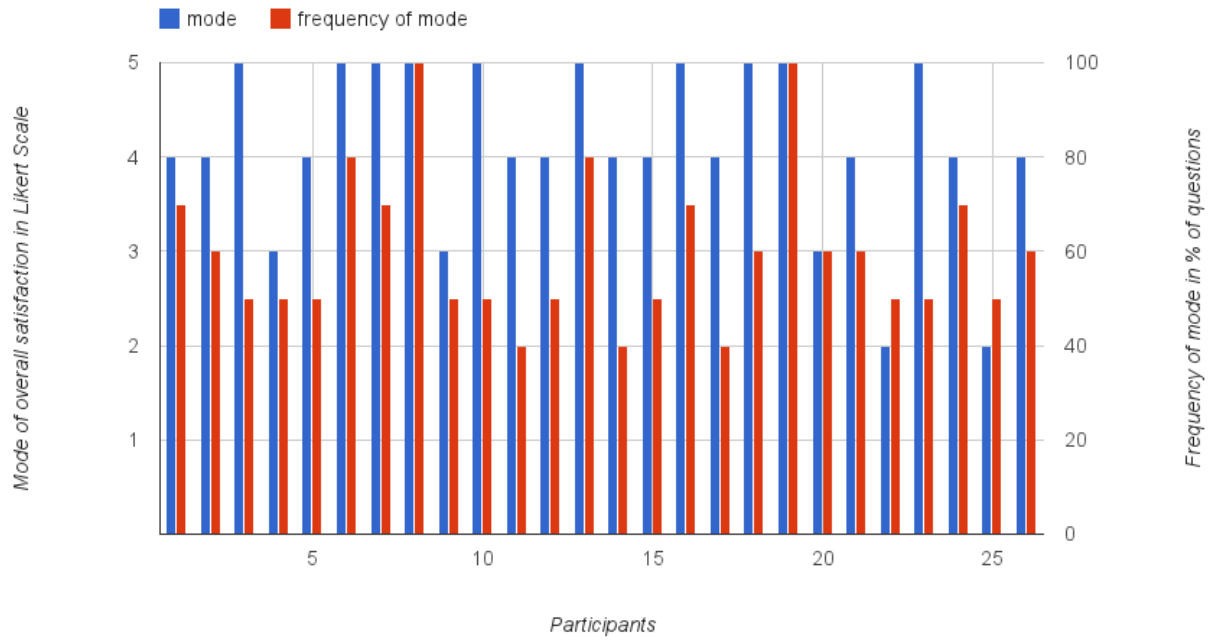

Figure 13: Mode and mode frequency values for each participant over the global satisfaction questions in Likert scale and % of questions.

Both are shown in Figure 13.

The run time of each part and the time consumption of a user test is visualized in Figure 14. The left hand side shows that the average total runtime of one user test was about 65 minutes with the colors indicating the connection of the part to the introduction in blue, the surveys in green, the 2D prototype in orange and the 3D prototype in yellow. As expected, the most time consuming parts include the 3D prototype taking almost 40% of the total runtime, shown on the right side of Figure 14.

In the following sections a detailed description of the results of each subsystem prototype evaluation is given.

3.1 2D Image/article search subsystem

The results of the 2D Image Search subsystem evaluation are presented in this section. The average time for putting the first result into the Tray tool (an action that declared that a result is relevant, according to the task description) for the 2D image task was 102 seconds. For the task using the Semantic Image Search it was 86 seconds. The article search task took on average 159 seconds to find the first relevant result. The success rates for the 2D image task, semantic image task and article search were $100\%, 100\%$ and 85% respectively. A task was considered successful if at least one relevant result was found within the time limit of the task. The respective average numbers of found results were $5.2, 4.5$ and 3.0.

The user satisfaction over specific aspects of the system collected from the questionnaires is given in Figure 15 in comparison to the results of the previous prototype. It is graded on a

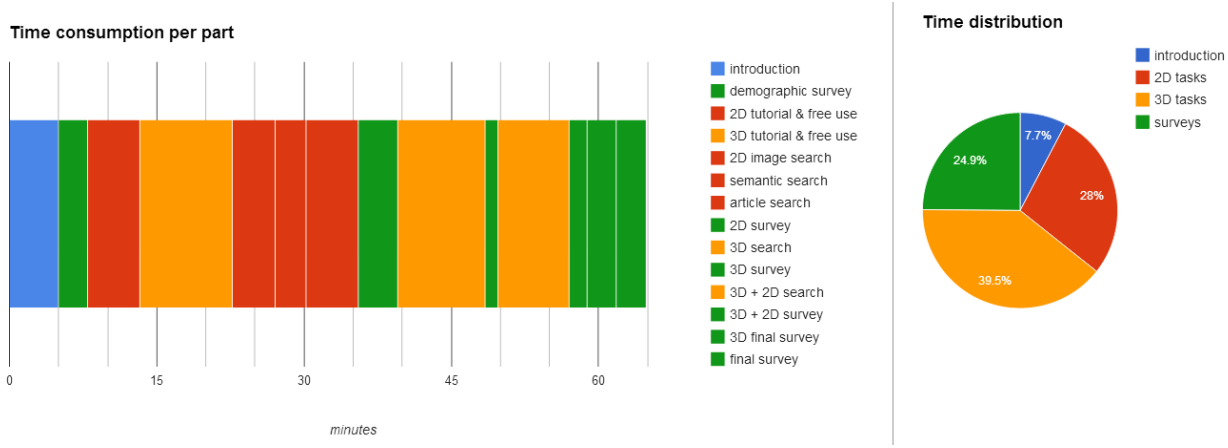

Figure 14: On the left side the run time of one user test. On the right side the time distribution between introduction, tasks and surveys.

Likert scale where 1 is strongly negative and 5 is strongly positive. The results satisfaction and presentation increased by a scale value of 2, to 4 and 5 respectively. Also with the response time and error correction support the users were more satisfied with an increase from 4 to 5. To assess the global satisfaction of each participant, the mode and its frequency over the questions is given in Figure 16. 24 out of 26 participants gave a mode above average (4 or better).17 out of these 24 gave that grade with frequency $\geq 50\%$, which means that they gave this answer in more than half the question.

Below follow the comments about the 2D Image Search subsystem, obtained by the open questions and the observations of the task performance. They are organized in three main categories, according to the part of the system that they refer to and their type: Frontend, Backend and Bugs. The Frontend list contains comments regarding the interface of the system. The Backend list contains comments regarding the tool functionality (e.g. search results, modality filtering etc.). The Bugs list contains irregularities causing erroneous behavior of the system. A number at the end of the line indicates the number of participants having this comment, while general observations by the observers do not have these numbers.

Frontend

- Drag and drop, participants expected to directly drag and drop items without selecting them. - all
- More clear workflow structure, current order: top left, top/mid right, mid/bottom left, bottom right.
- Windowing function in 2D needs to be more sensible.
- After making a selection in a drop-down menu, continuing is not intuitive (it overlies the search button, users have to click somewhere else or exactly at the top of the drop down menu to close it).
- Participants would like to use keyboard to jump to the words starting with specific letter in the drop-down menu, scrolling through is frustrating. - 5

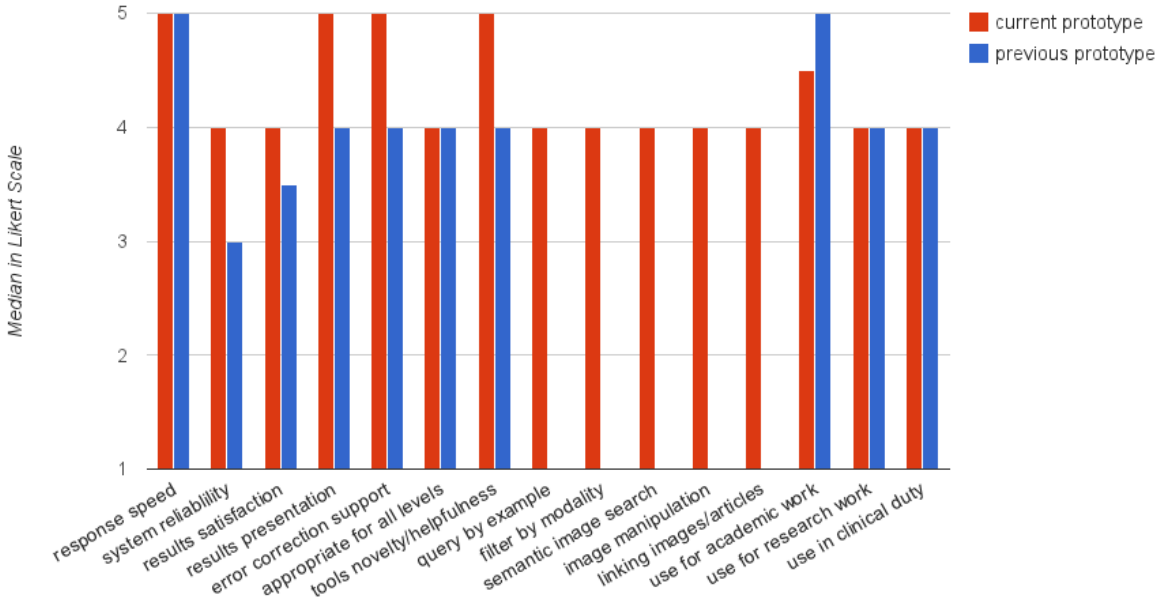

Figure 15: User satisfaction survey on the 2D/article search subsystem compared to the previous prototype.

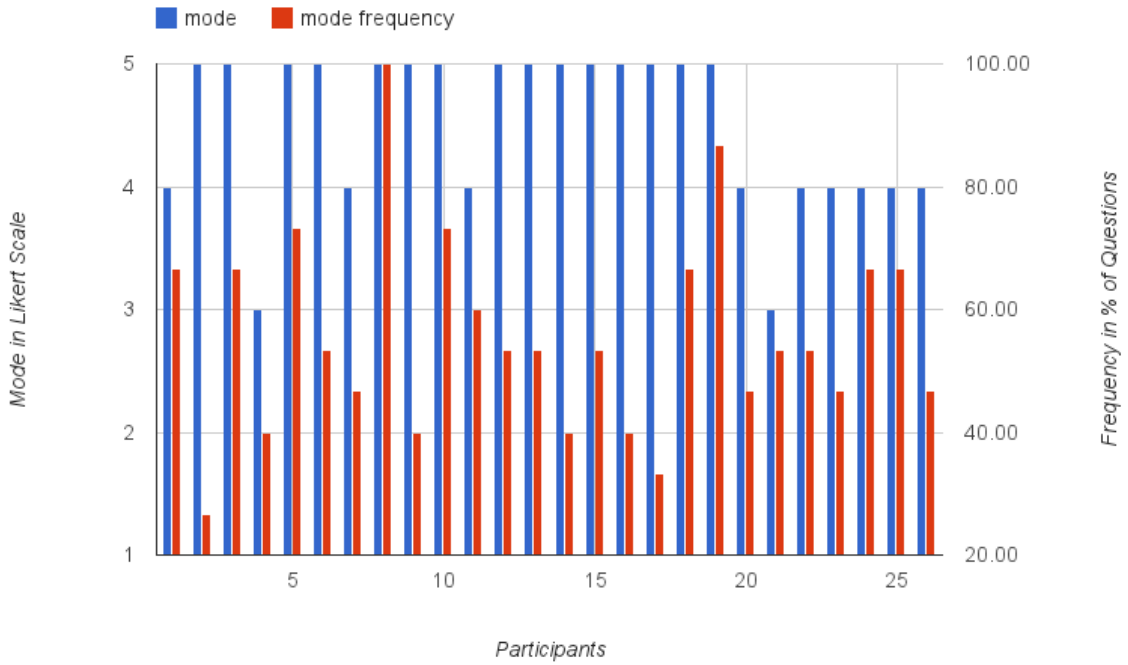

Figure 16: Mode and mode frequency of overall satisfaction on the 2D/article search subsystem per participant.

- Extraction of semantic terms should be automatic when typing and be given as a suggestion -2
- Result images should be larger, show e.g. 2 rows with 3 images each. - 5
- In the detail window participants always have to scroll down to view (larger) images. - 5
- Participants have a hard time selecting from the drop down menu (semantic imge tool), they have to click several times until its marked as selected, maybe the click area is too small? - 5
- Difficulty finding the modality search filter - 2
- Sorting results by modality was found to be very helpful - 6

Backend

- When selecting multiple modalities the participants expected that OR is used (instead of AND, where no results showed up) - 6
- Participants expected more results to show up, especially for the first task with only about 12 osteoporosis images. - 8
- Results are too rare/specialized and would not help during clinical routine, would like more general results - 5
- Include radiographics as additional source - 1
- Users are irritated when the exact same images show up that were marked as non-relevant. - 2
- Semantic Image Search missing anatomy brain as option, should accept terms that are not listed. - 2
- When searching for a specific modality, images from other modalities show up (e.g. CT $\rightarrow$ x-ray results). - 4
- Negative results should affect the modality filtering - 3

Bugs

- When adding multiple images the placement order of dragged 2D images seems to be inconsistent.
- Using right click, add as non-relevant in Semantic Image Search causes the system to switch back to the 2D Image Search.
- The Select Sources tab sometimes does not show up when trying to open it in the Image Search Perspective.

- Query with only non-relevant images is not working. - 4
- Using images only and no text does not give any results. User does not know why (text is required, e.g. do not allow search when textbox is empty!) - 1
- For some time no images were returned for a query, but it worked again after about a minute. - 2
- Images did not load immediately with search - 1
- The image in the search query was on the far left and not visible (only its boundary), and there was no scroll option.
- Clicking image search after article search enables additional undesired resources in the image search.
- Drag and drop from article detail view was not working, had to press add to query - 3
- Personal library image was zoomed in and caused the system to slow down - 10
- cannot use keyboard arrows to move between personal library items. - 1
- In the personal library participants do not realize that they are looking at an image with a different tag. - 3
- Personal library was not available. - 7
- Reset content does not remove tabs from DETAIL view.
- The result tab showed up in the query tab.

3.2 3D Image search subsystem

The design of the 3D prototype started off by creating six different mock-up designs of the front end, which were then discussed in interviews with physicians [6]. Using this feedback the first prototype was developed and the first tests with users were run on the alpha version. This helped to identify the initial bugs and to refine the user test protocol that was used for the first user tests. Their focus was on the evaluation of the work flow and the user interaction and the feedback was again used to improve the work flow of the prototype. As the performance and quality of searching for volumes without a ROI (anatomy search) was satisfied in the first user tests, the final user tests focus on the quality of the results and speed of the prototype concerning the pathology retrieval, as previously the time to get results using a ROI was about 1 minute. Requests from the physicians that were included in the prototype are image manipulation options as well as the linking of both prototypes.

The 3D prototype was evaluated by 25 participants, as during one user test the 3D system became unavailable. The user satisfaction over specific aspects of the system collected from the questionnaires is given in Figure 17, where the current results are shown in red and the results from the previous user test using the first prototype in blue. The biggest difference is the increase of 2 scale points in terms of response speed and result satisfaction, reaching median

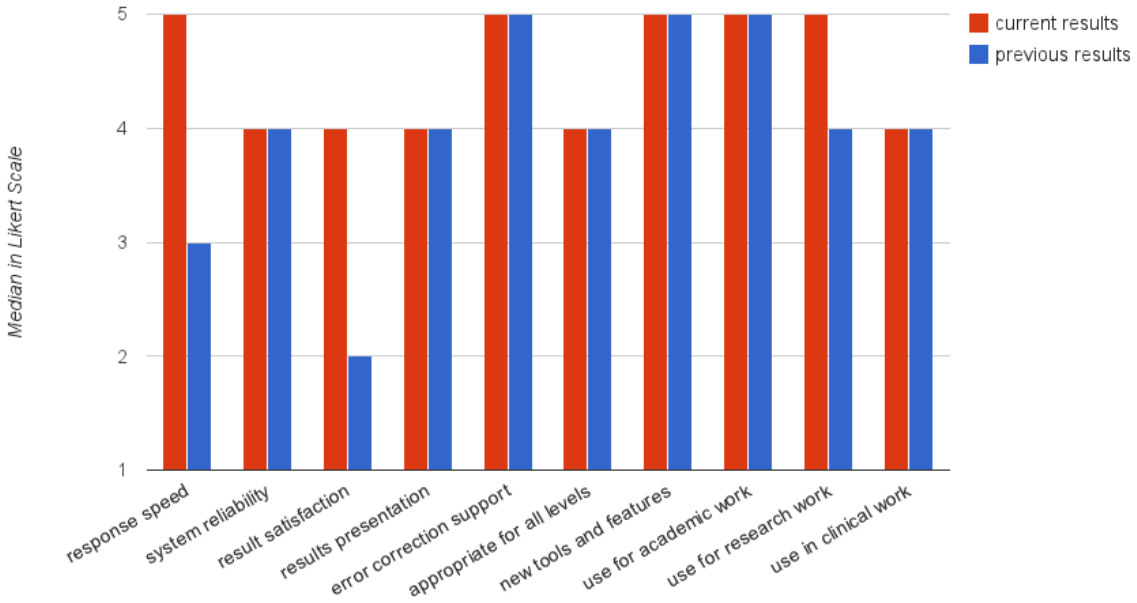

Figure 17: Results of the user satisfaction survey on the 3D system.

values of 5 and 4 respectively. The use of the prototype for research work also increased from 4 to 5, while the other results did not change. To assess the global satisfaction of each participant, the mode over the questions together with its frequency is given in Figure 18. 23 out of 25 participants gave a mode above average (4 or better). 17 out of these 23 gave that grade with frequency $\geq 50\%$, which means that they gave this answer in more than half the question.

The average time until the first result that satisfies the user is found is 259 seconds, while the users spent an average of 448 seconds in total for a 3D task. In 86% of the searches the users found satisfying results.

Below follow the user comments on the 3D Image Search subsystem, obtained by the open questions and the observer of the tasks. The organization is the same as Section 3.1.

Front-end

- Make images larger, fill the unused space, especially in the ROI marking view (volume-only-view). - 5
- Cannot stop / cancel loading volume in the result view - 4.
- When using the advanced filter option, also mark the matching term in the detail-view report - 2
- In the article detail view only mark the searched terms, not the rest - 2.
- Participants would like to use multiple filters in the result view (e.g. other patient + consensus + age) - 4

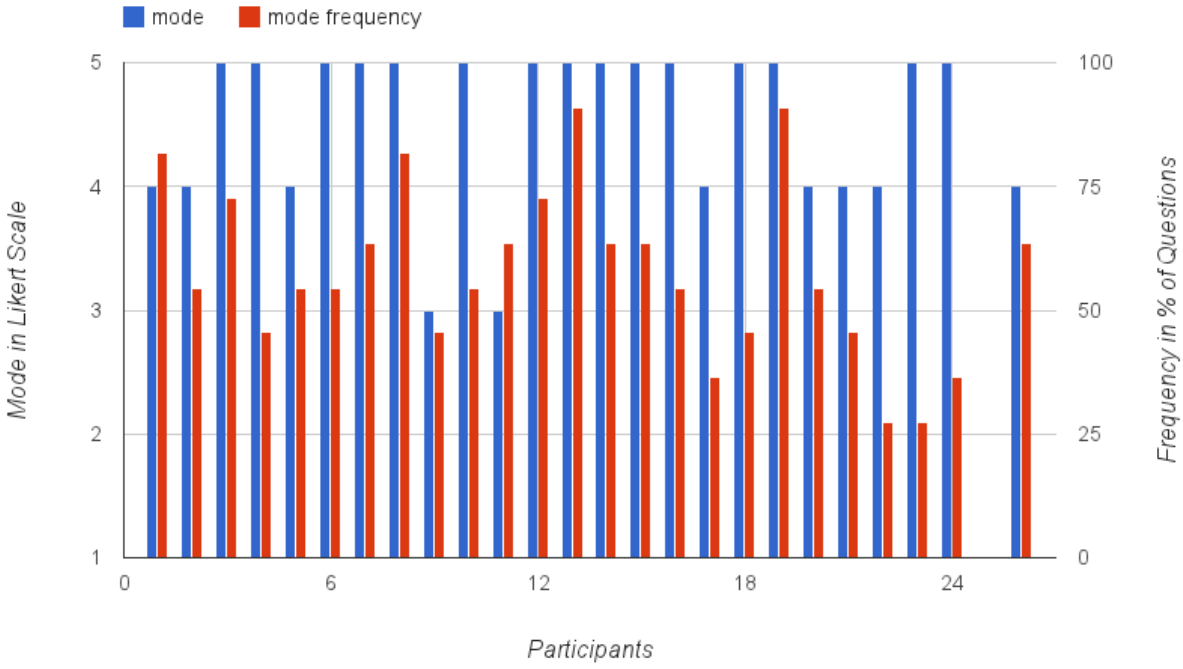

Figure 18: Mode and mode frequency of overall satisfaction per participant.

- Participants did not realize whether marking mode was enabled or not - 4
- Make the switching between perspectives more easy, 1 button instead of having to use menu. - 6
- There is no progress information for loading the volume. - 7
- There is no progress information about if the integrated 2D/Text-tabs will show up. - 6
- There is no information in the detail view if the volume is from the same patient or not that can be immediately noticed - 5
- Participants would like to be able to search directly from the marking view in index perspective
- Using the result view for marking ROIs is difficult, brightness/contrast is slower, ROI overview is missing - 4
- The slice of the query volume in the result view should be the slice where a ROI was marked in - 10
- User would like to modify drawn ROI by moving, changing - 3
- Users would expect to adjust window levels when moving mouse to the whole screen (out of the volume) - 8

- Zoom in/out and pan buttons should be in volume-only-view - 1
- Participants wanted to use double click to open volume, but this causes different behavior (changes perspective) and they got confused. - 9
- Use the windowing setting of the query volume for result list - 5
- Predefined/preset windowing options (bone, lung,) - 9.
- Show windows level details in numbers. - 1
- Participants tried to use window function on thumbnails, would be nice to have - 7
- It would be nice to have more slices in the result thumbnails - 2
- Filtering the results using the advanced filter is very useful - 5
- Less color in the text, only mark searched terms. - 2
- Color coding was helpful - 1
- Blue and green hard to see, red-green blind - 2
- Better representation and structuring of the report. - 6
- Search button was expected to be at top left due to common habit (current PACS system in Vienna) - 7
- In the result view the number of slices would be interesting. - 1
- Labeling of the 2D/Article/BigData tabs is confusing - 1
- Limit search to patient returns nothing if there are no other recordings, participant thought its not working. - 1

Backend

- Participants selected multiple pathologies the first time. - 3
- The system should be able to find more specific/higher resolution features, only large bullae but not smaller ones were detected.
- There are different demands of features and resolution based on modalities and organs.
- Location of the pathology in the lung important (e.g. upper lobe) - 2
- Only search within similar recordings, e.g. do not return MIPs. - 5
- All results (also text search) only anatomy specific, e.g. if searched in lung do not return rib fractures, aorta, - 2

- Text-search, automatic text query for 2D is not so good, using own words is very useful - 7
- Change order of detail view, s.t. results are in the middle, and detail on the right - 13.
- Having consensus of organs different from the anatomy where the ROI was in is strange - 2.
- Use DICOMs instead of 16bit JPEGs to improve the windowing function. - 2
- Exam protocol should be included in the result view.

Bugs

- Volume view different from radiologists, they look from the feet up (and not from the top down as implemented).
- Zooming is problematic, participants cannot use windowing and cannot mark ROIs, also zooming in is slowing down the system. - 5
- The loaded volume did not match the selected one. Updated a few times until the correct volume was finally shown. (assumable caused by clicking on a lot of volumes, without loading them) - 2
- When clicking/selecting an index image using a filter-field, right view does not update accordingly, cannot compare directly if it is the one selected in index view - 12
- Rapidly clicking onto several volumes in the index view made the system slow, loading volume did not work, loading volume overview on the right of index view did not work anymore - 1
- The same patient was not indicated as same patient, although the query volume was the same. - 4
- Prototype ran out of memory, no notification for the user, images just did not load. Had to restart application - 8

4 Discussion

A total number of 26 persons participated in the final user tests of the KHRESMOI radiology prototype. The majority of users were young and mid-age with ages between 20 and 40 years and were involved in various radiology specializations. The level of experience in radiology also spread the whole spectrum of 0 to $>$ 10 years. As we already realized at the previous user-tests, recruiting radiologists was a difficult task, as they usually are on a busy schedule. Nevertheless we were able to recruit 20 radiologists in Vienna for the user tests. More recruitment was found in hospitals not associated with the project, to get broader feedback from radiologists with different backgrounds. Thus, 2 participants were recruited in Germany, 2 in Greece and 2 in Geneva, resulting in the already mentioned total of 26 participants. All participants were

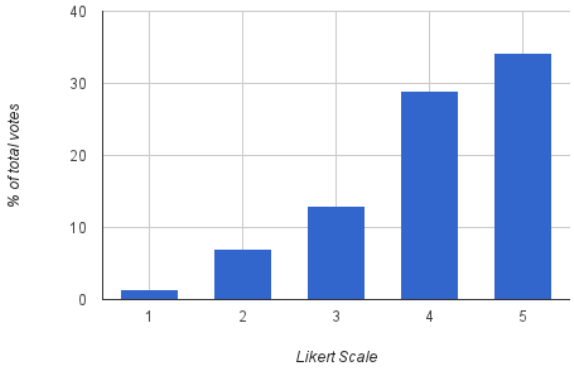

Figure 19: Normalized histogram of the general satisfaction questionnaire.

able to complete the 2D prototype tasks and all but one were also able to complete the 3D ones (due to network issues the 3D prototype became unavailable once). It is interesting to note that all the participants use the internet and internet search engines at more than once a day, either privately or in their job/education.

Main tendencies of the users satisfaction can be identified in Figure 12. Compared with the previous prototype the user satisfaction to use the system frequently increased from 4 to 5. The consistency was rated with a value of 4 which is also one value higher than previously. Users found that the system's design is straightforward and that is also easy to learn, without prior training. However, from the comments and the tester point of view a tutorial session would be beneficial. The main tendency of the users stayed strongly positive, with 10 persons giving a mode of 5 with frequency of at least 50% over the general satisfaction questions, which is also shown in the normalized histogram of all the given ratings shown in Figure 19.

In the following the results of the 2D and the 3D subsystems are discussed.

4.1 2D Image/article search subsystem

The average time taken to complete the image based tasks were 102 for the image search task and 86 seconds for the semantic image search task. This is an improvement over the 106 seconds reported in the first round of the user centered evaluation. The success rates of 100% for both image search tasks are higher than the reported 80.65% in year 2. Note that the times and percentages reported in the first evaluation round were already better than the self-assessments of the image search survey of the first year [7]. The time taken in information finding in the article search task is a little higher than year 2 (159 seconds compared to 150). However, the success rate of 85% is higher than the previous 78.95%. The mean number of relevant images found was slightly bigger than the year 2 prototype (5.2 and 4.5 compared to 4) even though the indexed dataset was significantly larger. This may have been caused by users stopping to search for additional results after the first relevant one was found, due to the time constraints of the participating physicians, albeit the instructions to search for as many as possible.

Regarding the user satisfaction, the most satisfactory aspects of the system seem to be the

response time, the results presentation and the ability to undo errors, all of them achieving a median of 5 on the Likert scale. The other parts of the system also obtained medians above average (4). The novelty of the tools is also recognized by most of the participants, getting a median of 5. The use of the system in academic work was found more slightly probable than use in research work or clinical duty achieving a median of 4.5 compared to 4 and 4 respectively.

The newly added features, such as modality filtering, semantic image search and image manipulation were satisfactory as well achieving mean grades of 4. From the feedback received in the open questions and the free discussion after the session, it was often mentioned that the semantic image search tool should use the auto-complete feature instead of drop-down menus for each field. Also regarding the modality filtering, even though it is very useful it should be more accurate, as several images where wrongly classified. There was a request that the relevance feedback should affect the modality filtering which is currently not taken into account. The interface was often found crowded or badly structured and a cleaner, bug–free and better structured version should be considered for a real world application.

Interesting suggestions for additions to the system were given, such as the ability to find similar articles when querying articles or images. The option to make the link of internal and external sources operate in both ways, was also mentioned. Currently the system offers the ability to have a search in internal sources extended to external sources. However, when an interesting case is found in the medical literature, searching for similar cases in the hospital records can also be of interest.

4.2 3D Image search subsystem

With a success rate of 86% most of the users found a satisfying result within an average time of 260 seconds. The differences to the Y2 prototype in Figure 17 show that the performance of the two worst performing points of the user tests could be drastically improved. The decrease in response time for a query decreased from 15 to 3 seconds and the loading time of the volume details from 45 to 17 seconds, resulting in a new satisfaction value of 5. Also the quality of the retrieval improved, now showing a satisfaction value of 4. The increase for use of the 3D prototype for research work from 4 to 5 may come from the integration of the 2D prototype, as these tabs give more information concerning research on a certain topic. While the overall response to the 3D prototype was very positive, the physicians had some additional suggestions that could help improve the work-flow and efficiency in using the prototype, including the following:

- In general the unused space should be used to increase the size of the thumbnails and volumes for improved visibility. In the ROI marking mode the volume could be shown in full-screen, as this is the only thing visible at this point.
- After pressing the search button, the query volume shown in the result view should by default show the slice where the ROI was marked. Manually searching for the correct slice is redundant and time consuming.
- The users also again mentioned that presets for brightness and contrast windows would be very useful. Also the set brightness and contrast window of the search volume should be automatically transferred to the result view and all the volumes shown there, this would

even remove the necessity of being able to change contrast and brightness for thumbnail images which is currently not implemented. The brightness and contrast manipulation mode should be available even when moving the mouse cursor out of the volume, as this was expected by the users.

- Another important comment of the users was to give them information about the progress of loading a volume and whether the integrated 2D search tabs will give any results or not.
- The coloring of the report satisfied some of the users, but dividing it into its basic structure (technical information, medical finding and diagnosis) would be more helpful to them.
- In terms of retrieval quality the users suggested to include information about the location of the pathology. Most important only return results for matching anatomies and second look at the local information of the pathology within the anatomy (e.g. search for pathology in the upper lobe of the lung). It would be also useful to limit the search to recordings with similar properties. Showing results from maximum intensity projections is not useful when searching for an normal Lung CT.
- The users also mentioned additional scenarios where the system could be useful. For example, if the head of radiology is currently not available, or if they would want to show students examples of certain pathologies (e.g. emphysema).
- During the tutorial and free use of the 3D prototype, participants searched for a variety of other pathologies. Most frequent and interesting ones were searches for granuloma, metastasis and pleura effusion. These could be specifically looked at in order to improve the quality of the retrieval engine.

5 Conclusion

This final round of user–centered evaluations concludes the development of the KHRESMOI Radiology prototype. The modifications that were made to the prototype after the first evaluation round, such as the addition of modality filtering, multi-modal relevance feedback, indexation of larger datasets and speedup plus the new 3D retrieval algorithm were found to be beneficial for the usability of the system. Retrieval times and success rates are improved compared to the year 2 evaluations, supporting the argument that the usability of KHRESMOI Radiology has improved by the modifications made after year 2. User satisfaction scores are also higher than the ones for evaluation of year 2, especially in aspects that had been lacking in performance, such as results quality, response speed and system consistency.

However, for the system to be usable in real life scenarios, additional modifications and improvements must be made. The modality filtering needs to achieve higher levels of accuracy and be more integrated with the relevance feedback functions. The semantic image search would be more user-friendly with the use of auto-completion features. The 3D retrieval needs to be more focused on the location of the ROI. Further research should be made in 3D retrieval when the marked ROI contains multiple pathologies. 2D image retrieval using ROIs could also improve the retrieval performance.

In conclusion, the KHRESMOI Radiology prototype is a tool that can potentially assist radiologists in information seeking scenarios for which the current tools are inadequate. Most of the participants that have tested the system have expressed intention of using a system like KHRESMOI in their academic, research and clinical work. Moreover the findings of this user study have revealed open challenges that need to be addressed to make such a system applicable to use in clinical environments.

6 References

[1] 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.

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

[3] 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.

[4] René Donner, Johannes Hofmanninger, Thomas Schlegl, Ljiljana Dolamic, Celia Boyer, Dimitrios Markonis, Henning Müller, and Georg Langs. Report on results of the second evaluation phase. Deliverable d2.7 of the khresmoi project, Medical University of Vienna, June 2014.

[5] Dimitrios Markonis, René Donner, Ljiljana Dolamic, Roger Schaer, Georg Langs, Célia Boyer, and Henning Müller. Report on and prototype of final image and analysis framework. Deliverable d2.6 of the khresmoi project, University of Applied Sciences Western Switzerland, February 2014.

[6] 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, University of Applied Sciences, Western Switzerland, 2013.

[7] 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.

[8] 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.

[9] 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.