Image Similarity Ranking of Focal Computed Tomography Liver Lesions Using a 2AFC Technique

Jessica Faruque[1], Sameer Antani[1], Rodney Long[1], Lauren Kim[2], George Thoma[1]

[1]Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health
[2] Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, Clinical Center, National Institutes of Health

ABSTRACT

Content-based image retrieval (CBIR) for radiological images has experienced massive growth over the past two decades, and shows great potential as a tool for use in precision medicine. A recurring challenge in CBIR evaluation has been in obtaining reference sets of images from human viewers of the system. Our work seeks to determine the feasibility of creating a reference set from images ranked by similarity from human viewers of the images. We obtained 2 sets each of 10 images of CT focal liver lesions from a database of open-access publications with and without markings showing the region containing the lesions, respectively. We created 2 sets of all 45 pair-wise combinations of the images, and displayed them to 10 volunteers, of which 2 had medical training. We used a Two-Alternative Forced Choice (2AFC) paradigm to obtain complete rankings of similarity levels in these image pairs. Analysis showed that inter-reader agreement for rankings ranged from Tau=0.21-0.69 (median=0.37) for the image pairs without any markings, and Tau=0.21-0.57 (median=0.33) for the image pairs with markings. A comparison of the regions of interests drawn by the study participants outlining the lesions in images without markings showed that participants tended to agree on images containing a single focal lesion of a single density, and inter-reader agreement for image rankings in which the regions of interest agree ranged from Tau=0.39-0.85 (median=0.58). These results show that the use of image ranking using 2AFC may be a feasible method for creating reference sets for CBIR system validation.

Keywords: Computed Tomography, image similarity, content-based image retrieval, liver, precision medicine

INTRODUCTION

Content-based image retrieval (CBIR) of visually similar radiological images has many potential applications, such as assisting in diagnosis in precision medicine, or providing medical training[1-6]. CBIR may be used in precision medicine to retrieve a customized set of image search results when the system is queried with radiological images from a specific patient. The interest in CBIR in medicine has grown over the last two decades. Many challenges, however, preclude the use of CBIR systems in medical training and decision support. One challenge is designing reference sets of visually similar images from large and diverse imaging datasets established by human viewers of the system who will be the ultimate users of a CBIR system. Another challenge is that while many sophisticated image processing and machine learning algorithms create CBIR systems, the utility of a particular CBIR system as a clinical decision support tool, training aid, and for other stated applications has yet to be definitively established[7-12].

While CBIR has existed for decades, medical image similarity perception for validating CBIR systems is a more recent innovation. Some of the current work in this topic includes asking readers to rate or rank similarity in mammographic images[13-19], however, mammographic studies are often focused on a binary classification or evaluation of benign or malignant diagnosis. Previous work with liver images, which have a variety of diagnoses, has shown moderate to high levels of inter-reader variability when readers were asked to provide numerical ratings of similarity between images [20-22]. At the same time, in tasks involving human judgment, image ranking has been shown to produce results with higher inter-reader agreement than the use of numerical ratings.

In this study, we seek to use image similarity ranking as a method of obtaining a reference set of image similarity for images of focal liver lesions. We use the 2-Alternative Forced Choice paradigm in which readers are presented with two

Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment,
edited by Craig K. Abbey, Matthew A. Kupinski, Proc. of SPIE Vol. 9787, 97870N
© 2016 SPIE · CCC code: 1605-7422/16/$18 · doi: 10.1117/12.2217364

pairs of images at a time and asked to select which pair of images is more similar. Ranking may be more efficient than providing similarity ratings since the person viewing the images has to make a yes-or-no decision rather than decide on a number, and we use a paradigm that minimizes the number of images that each reader views in order to obtain a complete ranking. We use participants with and without medical training, and we use both images which have markings indicating where a lesion is as well as images without any annotations. By performing this study, we determine the levels of inter-reader agreement obtained by the use of image ranking and if they are sufficient to develop reference sets in this manner for large datasets. These results will help develop more accurate reference standards for CBIR.

MATERIALS AND METHODS

Imaging Data and Image Selection Process: We used liver Computed Tomography (CT) images in this study, both because of the importance of making correct diagnoses of liver lesions, and also because of the broad spectrum of visual morphologies of liver lesions which make them an ideal model for observer training. We selected a total of 20 CT images of the liver for this study (Fig. 1) from the OPEN-i^{SM} multimodal biomedical image retrieval system (http://openi.nlm.nih.gov), which has been developed by the National Library of Medicine (NLM) and provides access to figures and images from NLM's PubMed Central repository.

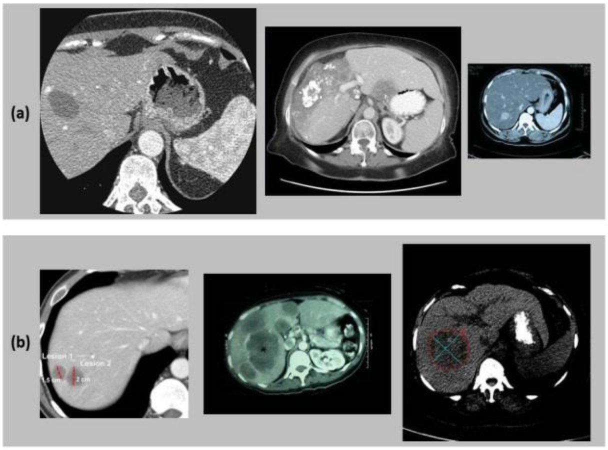


Figure 1: Examples of images used in the study (a) without arrows or markings, (b) with arrows or markings.

First, we obtained a set of search terms related to liver lesions from an ontology that was developed for liver disease. Next, we entered these search terms individually into the OPEN-i system, obtaining a total of 18272 image results, of which 5696 were unique. Out of these images, there were 281 and 397 axial CT images with and without arrows, respectively. A radiologist classified these images into three categories as good candidates for use in the study, not appropriate for use in the study, and possibly appropriate for use in the study. From the images classified as good candidates for the study, we selected 10 images with arrows and 10 images without arrows that had the highest ranking in the search results. Thus, we had 45 image pairs with arrows and 45 image pairs without arrows for this study (Fig. 2).

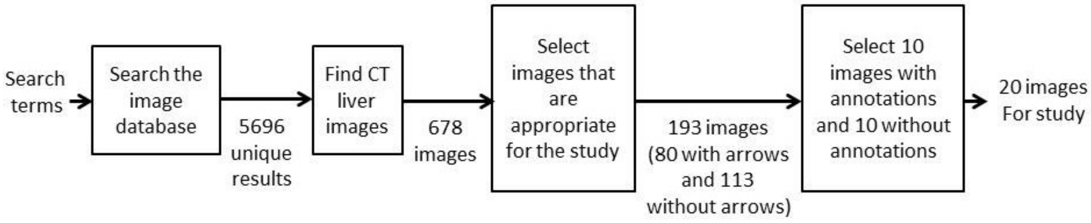
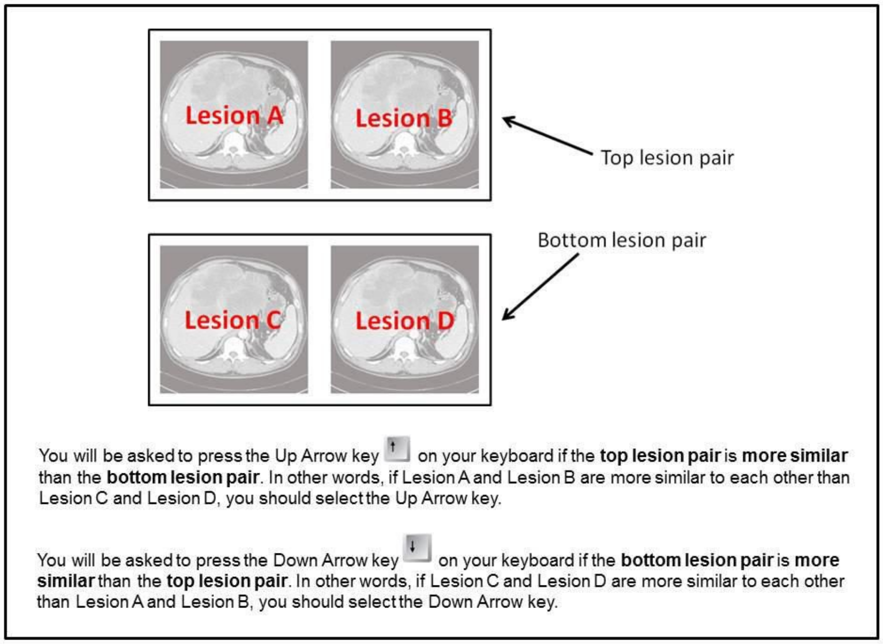

Search terms
Search the image database
5696 unique results
Find CT liver images
678 images
Select images that are appropriate for the study
193 images (80 with arrows and 113 without arrows)
Select 10 images with annotations and 10 without annotations
20 images For study

Figure 2: Image selection process for the study.

Study Participants and Participant Training: Institutional review board approval (IRB exemption #12909) was obtained for all of the data collection and analysis for this project. This study was completed by 10 participants, of which 2 had medical training, 3 did not have medical training but had seen or worked with medical images before, and 5 had no experience with medical images. A high percentage of our participants did not have medical training, however, there is evidence indicating that some types of training data collected from observers, such as image segmentation, may not necessarily require medical expertise[23].

All of the participants were provided with a tutorial prior to comparing similarity in the images. The tutorial first explained what axial CT images looked like, how to identify regions of the liver within the images, and how to identify lesions in the liver. Next, the tutorial showed a diagram of the task (Fig. 3). Finally, the tutorial ended with a brief seven-question, multiple-choice quiz to test the participants' knowledge of the material. Participants that marked erroneous answers engaged in a verbal discussion of the question to ensure that they understood the correct answers. For participants with medical training, most of the medical information in the tutorial appeared trivial, however, they were shown the same tutorial as the participants without medical training for consistency in this study.



You will be asked to press the Up Arrow key ↑ on your keyboard if the **top lesion pair** is **more similar** than the **bottom lesion pair**. In other words, if Lesion A and Lesion B are more similar to each other than Lesion C and Lesion D, you should select the Up Arrow key.

You will be asked to press the Down Arrow key ↓ on your keyboard if the **bottom lesion pair** is **more similar** than the **top lesion pair**. In other words, if Lesion C and Lesion D are more similar to each other than Lesion A and Lesion B, you should select the Down Arrow key.

Figure 3: Diagram of the study task in the tutorial provided to participants.

Image Pair Ranking Paradigm: The goal of the participants' task was to rank each set of 45 image pairs in order of how similar they are, with the most similar pair being the highest ranked and the least similar pair being the lowest ranked. To achieve this, we used a technique known as Two-Alternative Forced-Choice (2AFC), in which readers view two pairs of images at a time and select the pair of images that they consider to be the most similar. By repeatedly presenting two objects (in this case, two pairs of images) at a time, and obtaining their relative ranking, 2AFC eventually obtains a ranking of all the objects in a set. This technique is commonly used in tasks involving human perception and psychology and has also previously been used for the study of similarity in mammographic images[24]. In order to obtain a ranking of all 45 images in each set while minimizing the number of comparisons made by each participant, we selected future images to be presented as a function of the image pairs that the participants had already viewed and compared. The image pairs were in essence ranked using a merge sort algorithm, in which the human participants were making the comparisons at each step. Since the merge sort algorithm has complexity O(n log n) where n is the number of objects, the number of comparisons performed by a participant for each 45-image-pair set was approximately 171. To avoid bias, we randomized the order in which each set of images were presented to each reader, and all of the readers first ranked the set of image pairs without arrows prior to ranking the set with arrows. The images were randomized and presented in a graphical user interface developed in Matlab (R2014a, Natick, MA) (Fig. 4).

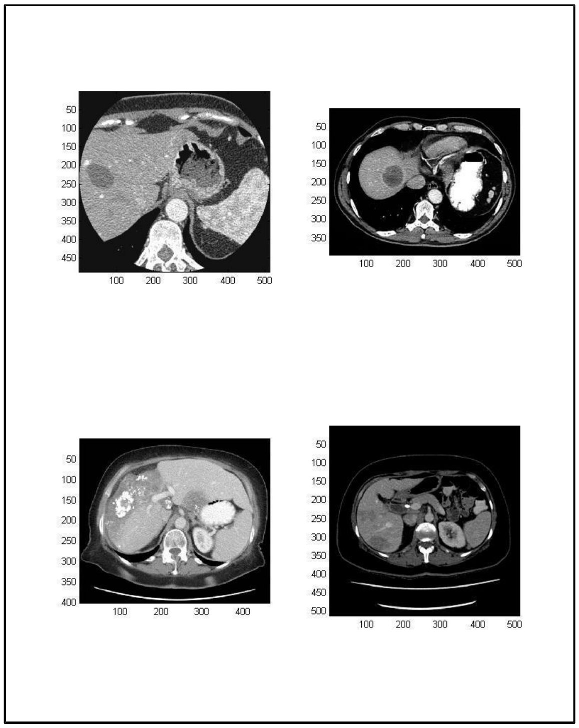


Figure 4: Screen capture of the Matlab interface developed. The images were scaled to fit each quadrant of the window.

Data Collection and Analysis: Prior to the study, we asked participants to fill out a questionnaire asking questions such as their training level and experience viewing medical images. During the study, we also collected data on the amount of time participants took to complete the study and the order in which the participants viewed the images, in addition to the similarity ranking scores. After the study, we asked readers if they had any questions or comments regarding their image interpretation in the study. Furthermore, we asked them to draw regions of interest (ROIs) in areas in the images without markings that they considered to be part of the lesion.

To analyze these data, we computed inter-reader agreement between each pair of readers by using Kendall's Tau metric, which ranges from 0 for no agreement to 1 for perfect agreement. We also determined which image pairs were consistently ranked high or low by all of the readers, and which image pairs had a good deal of discrepancy in their rankings. Next, we compared the amount of inter-reader agreement between readers with different experience levels, and the agreement levels of the rankings for the images with markings and the images without markings. Additionally, we compared agreement levels for images in which the readers drew either similar or different ROIs containing the lesion, and the amount of dispersion between readers for specific images.

RESULTS

In 6 of the images not containing arrows, the ROIs drawn by the participants seemed to have similar outlines among participants of both medical and non-medical backgrounds (Fig. 5). However, in the case of multiple potential lesions or multiple textures, the outlines selected by the participants varied widely.

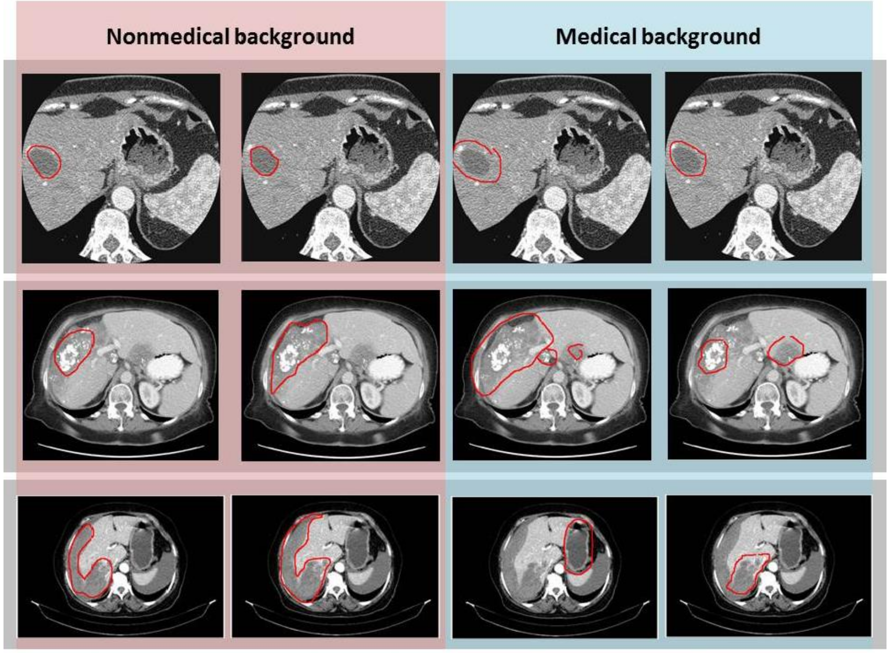


Figure 5: Example ROIs drawn by the participants for 3 of the images (without arrows or markings already drawn) in the database. The two participants on the left did not have medical training, whereas the two participants on the right did.

Analysis showed that the inter-reader agreement for rankings ranged from Tau=0.21-0.69 (median=0.37) for the image pairs without any markings, and Tau=0.21-0.57 (median=0.33) for the image pairs with markings (Fig. 6). Inter-reader agreement for image rankings in which the regions of interest agreed ranged from Tau=0.39-0.85 (median=0.58). We performed Wilcoxon signed-rank tests between the three distributions that indicated that the differences between these three sets of inter-reader agreement values were different (p<0.05).

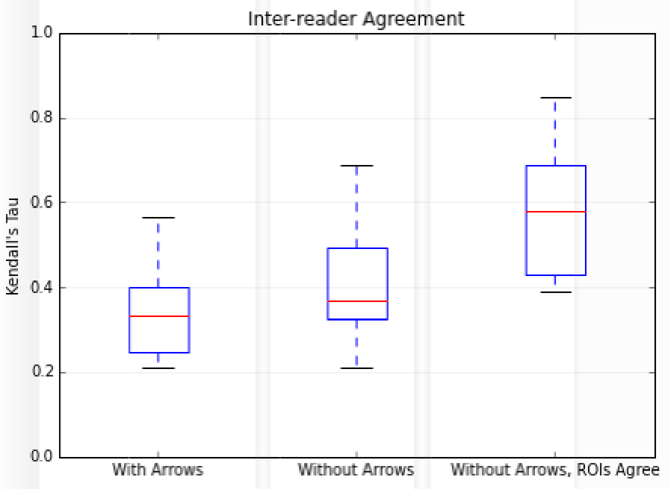


Figure 6: Inter-reader agreement for similarity between images with arrows, without arrows, and the subset without arrows in which the ROIs agree.

DISCUSSION

Our study involved both i ages with and without markings indicating where the lesions are. We found that lower inter-reader agreement occurred with the images with markings than without the markings. One reason that this may have occurred is because the images with markings varied in the types of markings, with some of them displaying a border around the lesion, and others displaying an arrow or asterisk to indicate where the lesions are. In the end, perhaps participants were more consistent in interpreting the images themselves rather than attempting to interpret what the markings mean.

The ROIs drawn by the participants seemed to be relatively consistent in images that contained one focal lesion of a single density. However, for images containing multiple lesions, or lesions with multiple densities, the outlines were much more variable. This may have contributed to the high inter-reader variability in the study, and may perhaps be prevented in the future with additional training and feedback to the participants before they complete the study tasks. Interestingly, some of the ROIs drawn by the participants with medical training (both had completed medical school) were less accurate than the participants without training; this may be because of the relatively little training in CT image interpretation provided to medical students.

We asked participants to rank the images using a 2AFC technique. Alternative techniques include asking readers to provide numerical similarity ratings between images, showing a triad of images and asking the reader to select which image is the least similar to the others, and other methods using multiple images presented at a time. Each of these techniques have advantages and disadvantages. For example, with numerical similarity ratings, we can obtain distributions of how participants rated images, which may be useful to understand how they view image similarity, however, methods that force all image pairs to be ranked provides equal spacing between all of the image pairs. The latter may be useful when combining readers' rankings, because differences in the distributions do not have to be accounted for.

In this study, we decided to use images from open-access publications. These images were pre-selected by the authors to be the ones to be used in the study. The advantages of using images from publications were that a large number of images were easy to procure and were already anonymized. However, these images had variable window and level, lacked consistency in the devices used to acquire the image, and had a variety of image sizes and cropping, which may have contributed to an increase in inter-reader variability. One limitation of this study is the relatively small number of images used. We had a total of 20 images and obtained rankings for 2 sets of 10 images. Readers could rank the two sets of 45 image pairs that we generated from these images in a reasonable amount of time, however, using larger datasets may create reader fatigue. For larger databases, obtaining a complete ranking these images may be difficult, and combining relative rankings of small datasets into a single similarity matrix may be necessary.

CONCLUSION

This study seeks to determine if image ranking is a feasible technique of developing reference sets for CBIR by asking participants to rank images of CT focal liver lesions obtained from open-access publications. Inter-reader agreement in participants' rankings of image pairs varied from low to moderate, with higher inter-reader agreement occurring when we excluded images in which participants disagreed on their interpretation of the ROI boundaries. The participants tended to agree more for ROIs that contained a single focal liver lesion. Future work involves studies that directly compare the effects of using images from publications to using standard formats such as DICOM, and performing these studies on larger datasets.

ACKNOWLEDGMENTS

The authors would like to thank the study participants for their time and participation in the study. This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

REFERENCES

[1] Muller H, Michoux N, Bandon D, et al. A review of content-based image retrieval systems in medical applications: clinical benefits and future directions. International Journal of Medical Informatics. 2004;73:1-23.

[2] Muller H, Rosset A, Garcia A, et al. Benefits of content-based visual data access in radiology. Radiographics. 2005;25:849-858.

[3] Eakins JP. Towards intelligent image retrieval. Pattern Recognition. 2002;35:3-14.

[4] Datta R, Joshi D, Li J, et al. Image retrieval: ideas, influences, and trends of the new age. ACM Computing Survey. 2008;40:5:1-5:60.

[5] Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized Medical Imaging and Graphics. 2007;31:198-211.

[6]Akgul DL, Napel S, Beaulieu CF, et al. Content-based image retrieval in radiology: current status and future directions. Journal of Digital Imaging. 2011;24:208-222.

[7] Muller H, Muller W, Squire D, et al. Performance evaluation in content-based image retrieval: overview and proposals. Pattern Recognition Letters. 2001;22:593-601.

[8] Krupinski EA, Berbaum KS. The medical image perception society update on key issues for image perception research. Radiology. 2009;253:230-233.

[9] Krupinski EA. The role of perception in imaging: past and future. Seminars in Nuclear Medicine. 2011;41:392 - 400.

[10] Manning DJ, Gale A, Krupinski EA. Perception research in medical imaging. British Journal of Radiology. 2005;78:683-685.

[11] Beam CA, Krupinski EA, Kundel HL, et al. The place of medical image perception in 21st-century health care. Journal of the American College of Radiology. 2006;3:409-412.

[12] Smeulders AWM, Worring M, Santini S, et al. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000;22:1349-1380.

[13] Muramatsu C, Li Q, Schmidt R, et al. Investigation of psychophysical similarity measures for selection of similar images in the diagnosis of clustered microcalcifications on mammograms. Medical Physics. 2008;35:5695-5702.

[13] Muramatsu C, Li Q, Schmidt R, et al. Experimental determination of subjective similarity for pairs of clustered microcalcifications on mammograms: observer study results. Medical Physics. 2006;33:3460-8.

[15] Muramatsu C, Li Q, Suzuki K, et al. Investigation of psychophysical measure for evaluation of similar images for mammographic masses: Preliminary results. Medical Physics. 2005;32:2295-2304.

[16] Muramatsu C, Li Q, Schmidt RA, et al. Determination of similarity measures for pairs of mass lesions on mammograms by use of BI-RADS lesion descriptors and image features. Academic Radiology. 2009;16:443-449.

[17] Li Q, Li F, Shiraishi J, et al. Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules. Medical Physics. 2003;30:2584-2593.

[18] Muramatsu C, Li Q, Schmidt R, et al. Investigation of psychophysical similarity measures for selection of similar images in the diagnosis of clustered microcalcifications on mammograms. Medical Physics. 2008;35:5695-5702.

[19] Wang J, Jing H, Wernick MN, Nishikawa RM, Yang Y. Analysis of perceived similarity between pairs of microcalcification clusters in mammograms. Med Phys. 2014 May;41(5):051904.

[20] J. Faruque et al., “A scalable reference standard of visual similarity for a content-based image retrieval system,” IEEE Healthcare Inf. Imaging Syst. Biol. 158–165 (2011).

[21] J. Faruque et al., “Modeling perceptual similarity measures in CT images of focal liver lesions,” J. Digit. Imaging 26(4), 714–720 (2013).

[22] J. Faruque et al., “Content-based image retrieval in radiology: analysis of variability in human perception of similarity.” SPIE Journal of Medical Imaging 2(2), 025501 (2015).

[23] Maier-Hein L, Mersmann S, Kondermann D, et al. Can Masses of Non-Experts Train Highly Accurate Image Classifiers? MICCAI 2014 Lecture Notes in Computer Science 2014, 438-445.

[24] Nakayama R, Abe H, Shiraishi J, et al. Evaluation of objective similarity measures for selecting similar images of mammographic lesions. Journal of Digital Imaging. 2011;24:75-85.