Effect of Reference Image Retrieval on Breast Mass Classification Performance: ROC Analysis

Chisako Muramatsu[1], Tokiko Endo[2,3], Mikinao Oiwa[2], Misaki Shiraiwa[3],
Kunio Doi[4,5], Hiroshi Fujita[1]

[1] Department of Intelligent Image Information, Gifu University, Gifu, Japan
{chisa, fujita}@fjt.info.gifu-u.ac.jp
[2] Department of Radiology, Nagoya Medical Center, Nagoya, Japan
endot@nnh.hosp.go.jp
[3] Department of Advanced Diagnosis, Nagoya Medical Center, Nagoya, Japan
[4] Gunma Prefectural Collage of Health Sciences, Maehashi, Japan
[5] Department of Radiology, The University of Chicago, Chicago, USA
k-doi@uchicago.edu

Abstract. Retrieval of similar cases with the diagnostic and therapeutic results as a reference may be useful in differential diagnosis of abnormalities. Image retrieval method for breast masses on mammograms has been investigated in our previous study, and the result indicated the potential advantage of a machine learning technique using sample cases with experts‘ subjective similarity data. In this study, the effect of presenting reference images to observers‘ ablity to distinguish between benign and malignant masses was investigated. Eleven physicians and 11 radiological technologists evaluated 98 masses and recorded their confidence of a lesion being malignant without and with reference images. The areas under the receiver operating characteristic curves improved from 0.926 to 0.938 (p=0.17) and from 0.895 to 0.928 (p=0.004) for the physician and technologist groups, respectively. The results indicate that reference images may be useful for diagnosis of breast masses.

Keywords: similar image retrieval; breast masses; digital mammograms; differential diagnosis; observer study

1 Introduction

Breast cancer is the most frequently diagnosed cancer in women in the U.S., some European countries, and Japan [1-3]. To reduce the number of death from breast cancer and to improve patients‘ quality of life, early detection and proper treatment are important. Periodic screening with mammography is considered effective for the early detection for women with normal risk [4-6].

When a new lesion is found on mammograms, it is generally evaluated with other image modalities. Even in such situation, it is beneficial to thoroughly evaluate mammographic findings to compare with those in other modalities. However, it is not easy to make differential diagnosis of lesions on mammograms. It has been suggested that computer-aided diagnosis (CAD) that provide the likelihood of malignancy of lesions may be useful in improving radiologists‘ diagnostic accuracies in the observer performance studies [7-9]. On the other hand, numeric guide may not be sufficient to some radiologists. Compared to computer-aided detection in which radiologists are prompted the suspicious areas on mammograms, the likelihood generally lacks in providing evidence to radiologists. Since radiologists‘ diagnostic ablity is based on experience, retrieval of similar images as a reference may be beneficial in providing supplemental information.

Several research groups have investigated the methods for automatic selection of similar images on mammograms and breast ultrasound images [10-19]. In earlier studies, the selection was based on the simple distance measures in the feature space [11], [13-14], [17]. In more recent studies, machine learning methods using samples with subjective similarity data were investigated, and the results were evaluated subjectively [15-16]. Other groups proposed specific decision making methods such as a 2-step selection method [18] and decision tree methods [19].

We have previouly investigated a similarity determination method using an artificial neural network (ANN), in which feature vectors of pairs of masses and the corresponding average subjective similarity ratings by experts were employed as input data and teacher, respectively [15]. In our recent study, a new method using multidimensional scaling (MDS) was introduced for understanding subjective similarity relationship between masses with different pathologies and for visualizing the subjective similarity space [20]. The objective similarity measure was determined on the basis of the distance in reconstructed subjective space using linear regression model. By leave-one-out cross validation test, the result indicated the usefulness of the MDS-based method.

In the present study, instead of applying ANN directly to estimate subjective similarity rating, MDS was first employed to map each image in subjective space, and the configured space was modeled by ANN [21]. After proper weights were determined, test cases were mapped in the modeled space, and the similarity between masses were determined by the distance in the similarity space. Using the proposed method, the effect of retrieved images on the readers‘ abilities to distinguish between benign and malignant masses was investigated in an observer performance study.

2 Materials and Methods

2.1 Mass Database

Digital mammograms used in this study were obtained at the National Hospital Organization, Nagoya Medical Center, Nagoya, Japan. The study protocol was approved by the institutional review board. The images were obtained with three digital sytems, including phase contrast mammography (PCM) system (Mermaid or Pureview, Konica Minolta Holdings, Inc.), direct conversion digital mammography system (Amulet, Fujifilm Corporation), and computed radiography systems (Mammomat 3000, Siemens, with C-Plate, Knoica Minolta, or Profect, Fujifilm). The original images have pixel sizes of 25 (PCM), 43.75 (C-Plate), or 50 (Amulet and

Profect) µm and grayscales of 10 (Profect), 12 (PCM and C-Plate), or 14 (Amulet) bits. For computational purposes, the pixel size and grayscale were unified to 50 µm and 10 bits, respectively.

Two radiologists reviewed the images and identified the masses by placing square regions of interest (ROIs) on the basis of the radiologic and pathologic reports. The ROIs were extracted from both craniocaudal (CC) and mediolateral oblique (MLO) views. When a lesion was partially cut off by the field of view, the ROI was excluded in this study. The size of the ROIs varied from 168 x 168 to 1888 x 1888. In this study, masses with 9 pathologic types were included: ductal carcinoma in situ (DCIS), invasive lobular carcinoma (ILC), mucinous carcinoma (MC), papillo-tubular carcinoma (PTC), scirrhous carcinoma (SC), solid-tubular carcinoma (STC), cyst, fibroadenoma (FA), and benign phyllodes tumor (BPT). PTC, SC, and STC are the subtypes of invasive ductal carcinomas, and invasive ductal carcinomas with unknown subcategories were not included in this study. The numbers of ROIs and lesions, and their mean effective diameters for the 9 types are listed in Table 1. The fractions of images obtained by different mammographic systems were 39% (Amulet), 23% (Profect), 21% (PCM), and 17% (C-Plate). All the malignant masses were confirmed by biopsy and/or surgery and benign masses were confirmed by biopsy or follow-up by mammography and ultrasonography.

Table 1. The numbers of ROIs and lesions and their mean effective diameters for 9 pathologic types used in this study

Pathologic type	Number of ROIs	Number of lesions	Mean effective diameter (mm)
Ductal carinoma in situ (DCIS)	14	10	24±13
Invasive lobular carcinoma (ILC)	12	7	34±9
Mucinous carcinoma (MC)	9	6	32±13
Papillotubular carcinoma (PTC)	39	21	27±12
Scirrhous carcinoma (SC)	69	40	35±13
Solid-tubular carcinoma	38	22	47±20
Cyst	99	63	25±18
Fibroadenoma (FA)	90	58	29±11
Benign phyllodes tumor	8	6	55±32
Total	378	233	28±16

2.2 Method for Similarity Determination

Determination of Subjective Similarity. For employing the MDS, in general, similarity (dissimilarity) data for all paired combinations of subjects must be obtained. To include a large variety of cases, but also retaining the number of comparisons by experts reasonably small, three masses from each of the 9 pathologic groups, thus a total of 27 masses, were sampled. As the result, subjective similarity ratings for all possible 351 pairs of masses were obtained independently by 8 physicians who have been certified for reading mammography by the Central Committee on Quality Control of Mammographic Screening in Japan. The details of the method and the analysis of the results have been described elsewhere [22]. Briefly, 2 ROIs for comparison was displayed in one monitor, and their entire views of the breast were displayed in another monitor. Each physician was asked to rate the similarity of the pair on a continuous scale from dissimilar (0.0) to similar (1.0) based on the overall impression for the shape, density and margin by taking into account the predicted pathology types. We asked them not to weigh on the size of the lesions, the surrounding normal tissue, and unrelated calcifications. The average subjective ratings were considered as the gold standard of similarity and used in MDS analysis.

Determination of Objective Similarity. Kruskal's nonmetric MDS in R programming language was employed. The configuration dimension was set to 3 in this study to reduce a risk of overtraining. For determination of ANN parameters, i.e., numbers of hidden units and iterations, a leave-one-out cross-validation was employed in a series of MDS analysis and configuration modeling. In this process, one ROI was removed, and MDS was applied to remaining 26 ROIs. Once subjective space was constructed, each dimension was modeled by ANN with 13 image features. These features were defined elsewhere [21]. The test ROI was then mapped to the reconstructed space by trained ANN. This process was repeated for all 27 ROIs. After all 27 ROIs were mapped, distances between all pairs of ROIs were determined, and they were converted to similarity measures by use of an exponential function. The adequacy of the model was evaluated by the correlation between the gold standard and the similarity measures.

2.3 Selection of Similar Images for Test Dataset

Usefulness of the similarity measure for selection of reference images was evaluated by precision, P, which is defined as

$$P = \frac{Number\ of\ images\ with\ matched\ pathology}{Number\ of\ retrieved\ images} \tag{1}$$

For the testing, MDS was applied to 27 ROIs, and the configured similarity space was modeled by ANN with the parameters selected by the leave-one-out regime. After excluding all the ROIs belonging to the same cases as the 27 ROIs, 324 ROIs were used for evaluation. For each test ROI, the most similar cases were selected from the 324 ROIs excluding the ones of the same case. If a query mass and the selected mass are both benign or both malignant, it was counted for matched pathology. The number of retrieved images was varied from 1 to 10 images, and the average precision was determined.

2.4 Observer Study

The effect of providing reference images in the diagnosis of breast masses was evaluated in the observer performance study. Ninety-eight cases, including 48 benign masses and 50 malignant masses, were randomly selected and included in the study. Eleven physicians who have been certified for reading mammograms and eleven radiological technologists who have been certified for mammography imaging with training of reading participated. In the reading session, a test ROI was displayed in one monitor, and the corresponding bilateral mammograms were displayed in another monitor. By clicking a button, images were switched between CC and MLO views on both monitors. After reviewing both views, observers were asked to mark their ratings on a continuous rating scale from definitely benign to definitely malignant. Subsequently, 5 reference images with their known pathologic types and the similarity map were shown, and the observers were asked again to mark their ratings. The results were evaluated by the multi-reader multi-case (MRMC) receiver operating characteristic (ROC) analysis [23] (MRMC software, the University of Chicago [24]). Figure 1 shows an observer study interface when reference images and similarity map were presented.

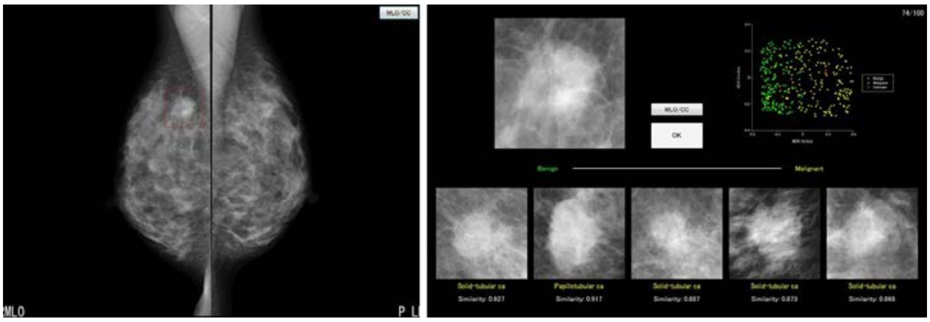

Fig. 1. Observer study interface with reference images. MLO views of the test case are shown in the left monitor, and the test ROI is shown in the right monitor with five reference images and their known pathologic types below and the similarity map on the top right.

3 Results

By the leave-one-out cross validation, similarity measures using MDS were determined for 351 pairs. The correlation coefficient between the gold standard and MDS-based similarity measures was 0.76, when the MDS dimension was 3. For comparison, the correlation coefficient between the gold standard and our previous ANN-based similarity measures for the 351 pairs was 0.68.

The method was applied to 324 test cases, and average precisions for the benign and malignant query images in retrieval of 1 to 10 images were determined and shown in Fig. 2. The average precisions were about 80% for both benign and malignant query images when 1 to 10 images were retrieved by the MDS-based measures. The result is relatively good as four of five reference images, on average, would be retrieved from the same benignity/malignancy groups as the test case. For comparison, the average precisions by the previous method were about 70% when 4 to 10 images were retrieved. The precision was slightly lower for the benign query images when small numbers of cases were retrieved. The proposed method was superior all across more than 100 retrieved images, although such large volume is impractical.

The observers' ability to distinguish between benign and malignant masses without and with reference images was evaluated in the observer performance study. By MRMC-ROC analysis, the average areas under the curves (AUCs) were 0.926 and 0.938 without and with reference images, respectively, for the physicians and 0.895 and 0.928, respectively, for the technologists. On average, AUCs for both groups were slightly improved. The difference was found to be statistically significant for the technologists (p=0.004), although we failed to find the statistically significant difference for the physicians (p=0.17).

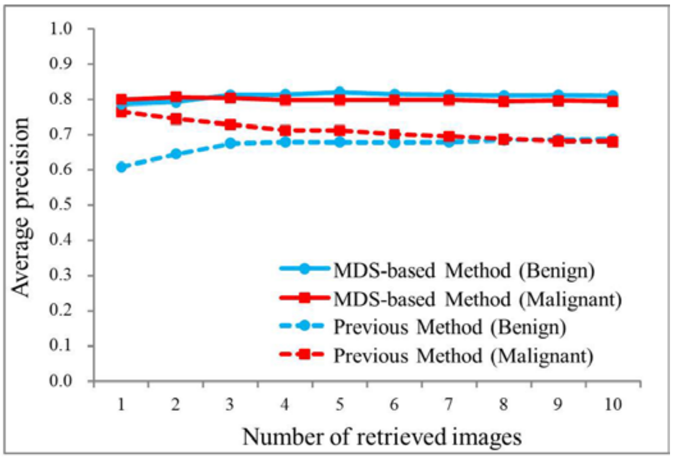

Fig. 2. Average precisions in retrieving pathology-matched reference images by the MDS-based similarity measures and previous ANN-based measures

4 Discussion and Conclusion

We have been investigating an effective image retrieval method to select images that are visually similar and useful in the point of view of diagnosis. Our new similarity measures are determined by applying MDS to the subjective similarity ratings obtained by experts for constructing a subjective similarity space and employing ANN to estimate the space with the image features. In this study, the proposed method was applied to the test cases, and the result was evaluated by precision in selecting pathology-matched reference images. When 1 to 10 images were retrieved, the majority of the cases (80%) were from the same pathologic group as the query images.

The effect of presenting reference images was evaluated in the observer study, in which observers' ability in distinguishing between benign and malignant masses was tested without and with the reference images. The average AUCs for both physician and technologist groups were slightly improved by showing reference images. Because the cases in this study were selected randomly and the observers in both groups were well trained, the AUCs were very high and the improvement was rather small. The technologists had the tendency to be slightly less confident at the initial reading and more likely influenced by the reference images. With the reference images, their average AUC was comparable to that of the physicians without reference images.

In conclusion, presentation of reference images may be useful in the diagnosis of breast masses on mammograms, especially for less experienced readers and slightly difficult cases. Our new similarity measures based on MDS may be effective in selecting useful reference images.

Acknowledgment

This study was partly supported by the Grants-in-Aid for Scientific Research for Young Scientists (No. 21791179) by Japan Society for the Promotion of Science and by a Grant-in-Aid for Scientific Research on Innovative Areas (No. 21103004), MEXT, Japan. Authors are grateful to following people for their contribution in this study and for participation in the observer study: K. Nishimura, T. Takahashi, T. Horiba, MD, M. Kato, MD, N. Kobayashi, MD., T. Morita, MD, M. Nishikawa MD, T. Niwa, MD, M. Sassa, MD, H. Yamamoto, MD, R. Suzuki, MD, Y. Hayashi, MD, N. Shinohara, J. Azuma, A. Imoto, J. Kikukawa, M. Yamada, Y. Suwabe, I. Morishita, A. Adachi, R. Kasugai, E. Katagiri, M. Tanaka, and E. Tukasaki.

References

1. Americal Cancer Society: Cancer Facts & Figures 2012. American Cancer Society, Atlanta (2012)
2. Ferlay, J., Autier, P., Boniol, M., Heanue, M., Colombet, M., Boyle, P.: Estimates of the cancer incidence and mortality in Europe in 2006. Ann. Oncol. 18, 581-592 (2007)
3. Matsuda, T., Marugame, T., Kamo, K.I., Katanoda, K., Ajiki, W., Sobue, T.: The Japan Cancer Surveillance Research Group. Cancer incidence and incidence rates in Japan in 2006: Based on data from 15 population-based cancer registries in the Monitoring of Cancer Incidence in Japan (MCIJ) Project. Jpn. J. Clin. Oncol. 42, 139-147 (2001)
4. Tabar, L., Fagerberg, G., Duffy, S.W., Day, N.E., Gad, A., Grontoft, O.: Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol. Clin. North Am. 30, 187-210 (1992)
5. Shapiro, S., Venet, W., Strax, P., Venet, L., Roeser, R.: Selection, follow-up, and analysis in the health insurance plan study: A randomized trial with breast cancer screening. J. Natl. Cancer Inst. Monogr. 67, 65-74 (1985)
6. Humphrey, L.L., Helfand, M., Chan, B.K.S., Woolf, S.H.: Breast cancer screening: A summary of the evidence for the U.S. preventive services task force. Annals. Internal Medicine. 137, E-347-367 (2002)
7. Chan, H.P., Sahiner, B., Roubidoux, M.A., Wilson, T.E., Adler, D.D., Paramagul, C., Newman, J.S., Sanjay-Gopal, S.: Improvement of radiologists' characterization of

mammographic masses by using computer-aided diagnosis: an ROC study. Radiology. 212, 817-827 (1999)

8. Huo, Z., Giger, M.L., Vyborny, C.J., Metz, C.E.: Breast cancer: effectiveness of computer-aided diagnosis – observer study with independent database of mammograms. Radiology 224, 560-568 (2002)
9. Jiang, Y., Nishikawa, R.M., Schmidt, R.A., Metz, C.E., Giger, M.L., Doi, K.: Improving breast cancer diagnosis with computer-aided diagnosis. Acad. Radiol. 6, 22-33 (1999)
10. Swett, H.A., Mutalik, P.G., Neklesa, V.P., Horvath, L., Lee, C., Richter, J., Tocino, I., Fisher, P.: Voice-activated retrieval of mammography reference images. J. Digit. Imaging. 11, 65-73 (1998)
11. Qi, H., Snyder, W.E.: Content-based image retrieval in picture archiving and communications systems. J. Digit. Imaging. 12, 81-83 (1999)
12. Sklansky, J., Tao, E.Y., Bazargan, M., Ornes, C.J., Murchison, R.C., Teklehaimanot, S.: Computer-aided, case-based diagnosis of mammographic regions of interest containing microcalcifications. Acad. Radiol. 7, 395-405 (2000)
13. Alto, H., Rangayyan R.M., Desautels, J.E.L.: Content-based retrieval and analysis of mammographic masses. J. Electron. Imaging. 14, 1-17 (2005)
14. Horsch, K., Giger, M.L., Vyborny, C.J., Lan, L., Mendelson, E.B., Hendrick, E.R.: Classification of breast lesions with multimodality computer-aided diagnosis: observer study results on an independent clinical data set. Radiology. 240, 357-368 (2006)
15. Muramatsu, C., Li, Q., Schmidt, R.A., Shiraishi, J., Doi, K.: Determination of similarity measures for pairs of mass lesions on mammograms by use of BI-RADS lesion descriptors and image features. Acad. Radiol. 16, 443-449 (2009)
16. Oh, J.H., Yang, Y., El-Naqa, I.: Adaptive learning for relevance feedback: application to digital mammography. Med. Phys 37, 4432-4444 (2010)
17. Xu, J., Faruque, J., Beaulieu, C.F., Rubin, D., Napel, S.: A comprehensive descriptor of shape: method and application to content-based retrieval of similar appearing lesions in medical images. J. Digit. Imaging. 25, 121-128 (2012)
18. Wang, X., Li, L., Liu, W. Xu, W., Lederman, D., Zheng, B.: An interactive system for computer-aided diagnosis of breast masses. J. Digit. Imaging. 25, 570-579 (2012)
19. Cho, H., Hadjiiski, L., Sahiner, B., Chan, H.P., Helvie, M., Paramagul, C., Nees, A.V.: A similarity study of content-based image retrieval system for breast cancer using decision tree. Med. Phys. 40, 012901-1-13 (2013)
20. Muramatsu, C., Nishimura, K., endo, T., Oiwa, M., Shiraiwa, M., Doi, K., Fujita, H.: Represenattion of lesion similarity by use of multidimensional scaling for breast masses on mammograms. J. Digit. Imaging. DOI: 10.1007/s10278-012-9569-0 (2013)
21. Nishimura, K., Muramatsu, C., Oiwa, M., Shiraiwa, M., Endo, T., Doi, K., Fujita, H.: Psychophysical similarity measure based on multi-dimensional scaling for retrieval of similar images of breast masses on mammograms. Proc. of SPIE Med. Imaging. (2013)
22. Muramatsu, C., Nishimura, K., Endo, T., Oiwa M., Shiraiwa, M., Doi, K., Fujita, H.: Correspondence among subjective and objective similarities and pathologic types of breast masses on digital mammography. In: Maidment, A.D.A., Bakic, P.R., Gavenonis, S. (eds.) IWDM 2012. LNCS, vol. 7361, pp. 450-457. Springer, Heidelberg (2012)
23. Dorfman, D.D., Berbaum. K.S., Metz. C.E.: Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest. Radiol. 27, 723-731 (1992)
24. Available at http://metz-roc.uchicago.edu/