Comparative Analysis of Data Collection Methods for Individualized Modeling of Radiologists' Visual Similarity Judgments in Mammograms

Georgia Tourassi, PhD, Hong-Jun Yoon, PhD, Songhua Xu, PhD, Garnetta Morin-Ducote, MD, Kathy Hudson, MD

Rationale and Objectives: We conducted an observer study to investigate how the data collection method affects the efficacy of modeling individual radiologists' judgments regarding the perceptual similarity of breast masses on mammograms.

Materials and Methods: Six observers of varying experience levels in breast imaging were recruited to assess the perceptual similarity of mammographic masses. The observers' subjective judgments were collected using (i) a rating method, (ii) a preference method, and (iii) a hybrid method combining rating and ranking. Personalized user models were developed with the collected data to predict observers' opinions. The relative efficacy of each data collection method was assessed based on the classification accuracy of the resulting user models.

Results: The average accuracy of the user models derived from data collected with the hybrid method was 55.5 ± 1.5%. The models were significantly more accurate ($P < .0005$) than those derived from the rating (45.3 ± 3.5%) and the preference (40.8 ± 5%) methods. On average, the rating data collection method was significantly faster than the other two methods ($P < .0001$). No time advantage was observed between the preference and the hybrid methods.

Conclusions: A hybrid method combining rating and ranking is an intuitive and efficient way for collecting subjective similarity judgments to model human perceptual opinions with a higher accuracy than other, more commonly used data collection methods.

Key Words: Breast imaging; mammography; observer variability; perception; visual similarity user modeling.

Collecting people's opinions regarding the visual similarity of images is a critical building block for developing content-based image retrieval (CBIR) systems (1,2). In recent years, CBIR has been proposed in clinical imaging to enhance clinical decision support and training systems with visually similar cases retrieved from a reference image library, thus emulating the evidence-based clinical paradigm (3,4). The reliability of the developed CBIR technology is closely tied to image similarity metrics that correlate highly with human perceptual opinions. The development and validation of such metrics depend on the number and diversity of medical images presented to radiologists during the data collection process, which is a rather time-consuming step. Moreover, CBIR systems often disregard human perception subjectivity and embrace a generalized modeling approach to reproduce the consensus opinion of several radiologists. Relevance feedback techniques (5–7) have been adopted in CBIR to capture human perception subjectivity by providing personalized fine-tuning of the image retrieval step. Still, this is work in progress in medical imaging (8–11).

The topic of perceptual subjectivity has been attracting attention in general (12–14) and for radiological applications in particular (15–22) with compounding evidence that the notion of visual similarity is highly subjective. Most studies use a rating-based data collection method wherein radiologists are asked to use a fixed rating scale (either continuous or discrete) to record their opinions regarding the similarity of image pairs. The rating method is well accepted in

Acad Radiol 2013; 20:1371–1380

From the Biomedical Science and Engineering Center, Oak Ridge National Laboratory, 1 Bethel Valley Road, P.O. Box 2008, Oak Ridge, TN 37831 (G.T., S.X., H.-J.Y.); and Department of Radiology, University of Tennessee Medical Center at Knoxville, Knoxville, TN (G.M.-D., K.H.). Received March 27, 2013; accepted August 6, 2013. **Address correspondence to:** G.T. e-mail: tourassig@ornl.gov

http://dx.doi.org/10.1016/j.acra.2013.08.002

psychometric and user studies (23). Among its main limitations are user inconsistencies in applying a numerical scale across multiple cases and personal biases due to internal cognitive processes and individual personality traits, which often result in people using only part of the rating scale (24–26). To the best of our knowledge, there has been only one study in radiology reporting relatively good agreement between continuous scoring versus discrete scoring, but the participating radiologists were more consistent using discrete scoring rather than continuous scoring (27).

Preference-based methods have been used in cognitive and affective modeling because they appear to overcome these limitations (28–30). Preference data collection eliminates the subjective notion of numerical scoring by asking a user to provide pairwise preference or the preferred order among many choices (30). However, user modeling with preference data can be rather limiting since it lacks quantitative information regarding pairwise differences. The preference data collection method has been used in radiology for demonstrating the feasibility of personalized user modeling of a radiologist's opinions (31) as well as to study the reproducibility of similarity ranking scores across multiple radiologists (32). A hybrid variation of the preference method was also explored as a means to derive a new "psychophysical" measure of visual similarity that captures the human perceptual judgment of image similarity (33). This study focused on group-based understanding of visual similarity, disregarding interobserver differences.

Extending these prior studies, we present our work, which brings attention to an issue mostly ignored in perceptual similarity studies—namely, the effect of the data collection method in deriving accurate and reliable user models for predicting individual opinions. Understanding whether the data collection method affects individualized modeling of radiologists' opinions regarding image similarity is an important step for building more-effective CBIR systems. Our study offers a systematic comparison among three methods: (i) a rating method, (ii) a preference method, and (iii) a hybrid method that combines the strengths of the preference and rating methods. The comparison is based on the predictive accuracy of personalized user models derived with data collected using each method respectively. Our overarching goal is to determine which data collection method facilitates the development of user models that can reliably capture subjective visual similarity across radiologists with different experience levels. Experiments were conducted for the same visual task as many of the studies cited earlier: similarity assessment of breast masses on mammograms.

MATERIALS AND METHODS

Image Database

Regions of interest (ROIs; 2.6 cm × 2.6 cm) containing biopsy-proved masses were obtained from the Lumisys volumes of the Digital Database of Screening Mammography (DDSM) (34). ROIs that (i) did not fully include the mass, (ii) were considered of poor image quality, and (iii) included calcifications that may influence radiologists' judgments were excluded from the study. Architectural distortions and focal asymmetries were also excluded. Forty ROIs depicting distinct mammographic masses of approximately similar size were randomly selected. The depicted masses represented the full range of shapes and margins according to the Breast Imaging-Reporting and Data System (BI-RADS) descriptors provided in DDSM. Of the 40 ROIs, 13 were extracted from left craniocaudal views, 10 from left mediolateral oblique views, 8 from right craniocaudal views, and 9 from right mediolateral oblique views. The final set included 26 malignant masses and 14 benign masses, shown in Figure 1.

Data Collection Method

Collection of observer data was done using three different study protocols. For each protocol, a different graphical user interface (GUI) was developed as an iPad (Apple Inc, Cupertino, CA) application. The design of a user-friendly, intuitive GUI on the iPad platform is essential for ensuring smooth operation without any unnecessary delays throughout the course of the study. We followed the guidelines for good GUI designs proposed by Stone et al (35). The following is a detailed description of each data collection protocol and the GUI implemented for the corresponding study protocol.

Rating method. The study participant is presented with a pair of masses, as shown in Figure 2. The participant is asked to provide a similarity score for the pair using a continuous scoring scale from 0 ("highly dissimilar") to 1 ("highly similar"). As mentioned earlier, this data collection method is the one used most often in radiology for human perception similarity and CBIR studies. The rating method relies on the assumption that radiologists use the scoring scale in a consistent manner throughout the study. However, it is a challenging expectation for a human to assign numerical scores in a coherent and reproducible manner.

Preference method. In this method, the study participant is presented with a triplet of masses (A, B, and C), as shown in Figure 3, and asked to identify the pair with the highest visual similarity. In contrast to the rating method, no numerical score is asked explicitly. Instead, the participant must make one of four possible choices; namely, select one of the three possible pairs of masses (A and B, A and C, or B and C) that appear visually most similar or report that no particular pair stands out as being most similar than the rest.

In general, the preference method presents an easier task to the study participants than the rating method, but that conclusion depends on the relative differences between the pairs. For example, intraobserver variability is unavoidable if the radiologist is asked to rank pairs of masses with similar BI-RADS characteristics (e.g., (31)). The main drawback of the preference method is the lack of absolute