Texture as a Diagnostic Signal in Mammograms

Yelda Semizer (yelda.semizer@rutgers.edu)
Department of Psychology, Rutgers University
152 Frelinghuysen Rd, Piscataway, NJ 08854

Melchi M. Michel (melchi.michel@rutgers.edu)
Department of Psychology, Rutgers University
152 Frelinghuysen Rd, Piscataway, NJ 08854

Karla K. Evans (karla.evans@york.ac.uk)
Department of Psychology, University of York
Heslington, York YO10 5DD, UK

Jeremy M. Wolfe (jwolfe@bwh.harvard.edu)
Department of Ophthalmology, Harvard Medical School
Department of Radiology, Harvard Medical School
Department of Surgery, Brigham and Women's Hospital
64 Sidney St. Suite. 170, Cambridge, MA 02139

Abstract

Radiologists can discriminate between normal and abnormal breast tissue at a glance, suggesting that radiologists might be using some "global signal" of abnormality. Our study investigated whether texture descriptions can be used to characterize the global signal of abnormality and whether radiologists use this information during interpretation. Synthetic images were generated using a texture synthesis algorithm trained on texture descriptions extracted from sections of mammograms. Radiologists completed a task that required rating the abnormality of briefly presented tissue sections. When the abnormal tissue had no visible lesion, radiologists seemed to use texture descriptions; performance was similar across real and synthesized tissue sections. However, when the abnormal tissue had a visible lesion, radiologists seemed to rely on additional mechanisms beyond the texture descriptions; performance increased for the real tissue sections. These findings suggest that radiologists can use texture descriptions as global signals of abnormality in interpretation of breast tissue.

Keywords: texture analysis; medical image perception; visual search; ROC curves, log likelihood ratios

Introduction

Human observers are able to obtain the "gist" of visual scenes within milliseconds (Friedman, 1979; Potter, 1976; Potter & Levy, 1969; Schyns & Oliva, 1994). Humans can rapidly categorize real world scenes as urban or natural (Greene & Oliva, 2009), or as indoor or outdoor (Fei-Fei, Iyer, Koch, & Perona, 2007). In a sense, humans are experts in categorizing natural scenes. Similarly, radiologists are experts at categorizing medical images as normal or abnormal (Evans, Georgian-Smith, Tambouret, Birdwell, & Wolfe, 2013; Evans, Haygood, Cooper, Culpan, & Wolfe, 2016; Kundel & Nodine, 1975). This ability allows them to extract necessary information to make quick and accurate diagnostic judgments.

Several models have been proposed to explain the search performance of radiologists, including "two-stage detection model" (Swensson, 1980), "two-pathway model" (Drew, Evans, Võ, Jacobson, & Wolfe, 2013; Wolfe, Võ, Evans, & Greene, 2011), and "global-focal search model" (Kundel, Nodine, Conant, & Weinstein, 2007; Kundel, Nodine, Thickman, & Toto, 1987; Nodine, Kundel, Lauver, & Toto, 1996). In general, these models propose two separate processes through which radiologists make diagnostic judgments based on medical images. First, radiologists rapidly form a global representation of images. They mark potential areas of abnormalities by comparing the present image to their template of normal and abnormal structures. Then, through further inspection of the previously flagged locations, they decide to categorize images as normal or abnormal.

Single glance studies suggest that expert radiologists can quickly extract global information from medical images, such as chest radiographs (Carmody, Nodine, & Kundel, 1980; Kundel & Nodine, 1975), computed tomography (CT) scans (Oestmann et al., 1988), and mammograms (Mugglestone, Gale, Cowley, & Wilson, 1995). There is evidence that radiologists can discriminate between normal and abnormal breast tissue after a very short period of exposure (250 ms) although they are unable to localize the site of abnormality (Evans et al., 2013). Recently, in a series of experiments, Evans et al. (2016) demonstrated a similar discrimination performance in cases where the briefly presented abnormal tissue (500 ms) had no visible lesion in it. To interpret this ability, authors suggested that radiologists might be using some signal of abnormality that is based on a global image statistic. Radiologists might be picking up this signal to make rapid and accurate diagnostic judgments. By interpreting thousands of images over several years, it is possible that radiologists become sensitive to such signal if it is present in the abnormal tissue.

Our study sought to characterize these global signals of abnormality as texture descriptions (i.e., a set of stationary spatial statistics) and to determine whether radiologists rely on such texture descriptions when discriminating between nor-

mal and abnormal breast tissue. To test this hypothesis, we generated synthetic images representing sections of breast tissue using a texture synthesis algorithm (Portilla & Simoncelli, 2000). The algorithm was trained on texture descriptions extracted from sections of mammograms confirmed via biopsy to be normal or abnormal. Synthetic images generated from a common model are physically different but have the same overall statistics (i.e., texture descriptions), so they should appear to the viewer as different sections of tissue from the same breast. Because the texture descriptions of the real and synthesized sections were identical, any global statistical signals of abnormality in the real sections were also present in the synthesized sections. We investigated performance of radiologists in a diagnostic task that required rating the abnormality of briefly presented tissue sections. Our results showed that radiologists seemed to rely on texture descriptions when the abnormal tissue did not have a visible lesion. However, radiologists seemed to use additional mechanisms beyond the texture description when the abnormal tissue had a visible lesion. Overall, our findings suggest that radiologists can use texture descriptions as global signals of abnormality in interpretation of breast tissue.

Method

Participants

A total of twenty-three radiologists participated in the study. Nineteen radiologists participated in the main experiment and eight (including four of the radiologists from the main experiment) participated in the control experiment. All participants had prior experience and training in reading mammograms.

Stimuli and Apparatus

Stimuli were images representing sections of breast tissue (256×256 pixels) which have been confirmed via biopsy to be normal or abnormal. Normal tissue sections were extracted from non-cancerous breasts. Abnormal tissue sections were extracted from cancerous breasts, and contained either a visible lesion (lesion-present) or no lesion (lesion-absent). In particular, abnormal "lesion-absent" sections were used to characterize the signal of abnormality in absence of a visible lesion in the tissue. Clinical breast density ratings confirmed that the normal and abnormal tissue sections were similar in breast density (all $r < |-0.08|$, all $p > 0.5$).

Synthetic textures were generated using a texture model (Portilla & Simoncelli, 2000) trained on the real tissue sections. Figure 1 shows examples of the real and synthesized tissue sections.

Stimuli were generated and presented using the Psychophysics Toolbox extensions (Brainard, 1997) in MATLAB (Mathworks) on a 24-in LCD monitor with a resolution of 1920×1080 pixels.

Procedure

Radiologists completed a task that required rating the abnormality of briefly presented tissue sections. At the start of each trial, a fixation cross was presented at the center of the screen. After the observer initiated the trial by pressing the start button, the image (i.e., a tissue section) appeared at the center of the screen for 500 ms, followed by a white noise mask for 500 ms. Then, a response screen with a slider was presented. The observer was asked to give a rating between 0 and 100 by sliding a bar to indicate the likelihood of recalling the patient. Once a decision has been made, the observer pressed a button to log their response. Figure 2 shows the timeline of a representative trial.

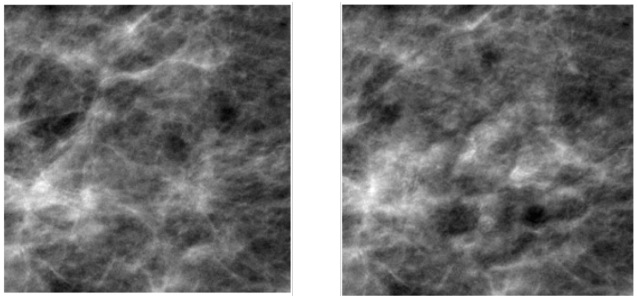

Figure 1: Example breast tissue sections, with real tissue on the left and synthesized tissue on the right.

The type of image (real or synthesized) and the type of tissue (normal or abnormal) were manipulated as within subject variables while the type of abnormality (lesion-present or lesion-absent) was manipulated as a between subject variable.

There was a total of 200 trials, divided equally between the real and synthesized conditions. For each condition, half of the trials included normal tissue sections while the other half included abnormal tissue sections, either lesion-present or lesion-absent. Trials were blocked by the type of image (real or synthesized) and the presentation order was randomized across observers. In each block, the presentation order of tissue type (normal or abnormal) was also randomized for each observer.

Breast density estimation In a control experiment, radiologists were asked to rate the density of briefly presented tissue sections. The procedure was similar to the task in the main experiment, except for the response screen. The response screen included four gray boxes numbered from 1 to 4, indicating BI-RADS breast density scale. Larger numbers indicated higher breast density. The observer was instructed to report the density of the tissue by navigating the boxes with button presses. The chosen box turned to red to indicate current choice of the observer. Once a decision has been made, the observer pressed a button to log their response.

Results

Performance was examined by constructing ROC curves. Because raw ratings tended to show *bimodal* distributions for normal and abnormal cases (see Figure 3), an optimal performance could not be determined using a single criterion. Therefore, in evaluating ROC curves, we used log likelihood

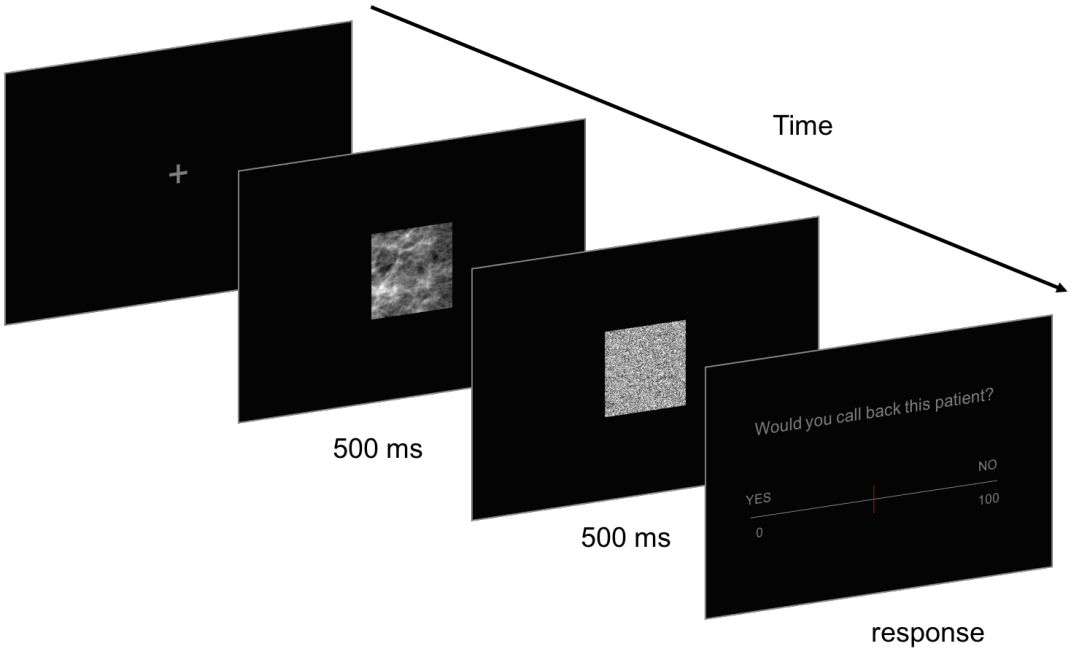

Figure 2: Timeline of a representative trial in the abnormality rating task.

ratios (LLRs) as the decision variable. LLRs were calculated using the formula,

$$LLR = \ln\left[\frac{p(x|abnormal)}{p(x|normal)}\right], \quad (1)$$

where x represents the raw ratings.

In order to prevent the possibility of over-fitting, we smoothed the data by fitting a Gaussian kernel with a bandwidth of 5. Figure 4 shows the resulting LLRs for the raw ratings shown in Figure 3. When the raw ratings were converted into LLRs, curves crossed at a single point (at LLR = 0). As a result, the optimal discrimination performance can be determined using a single criterion.

Using the smoothed data and LLRs as the decision variables, performance was characterized by computing the area under the curve (AUC).

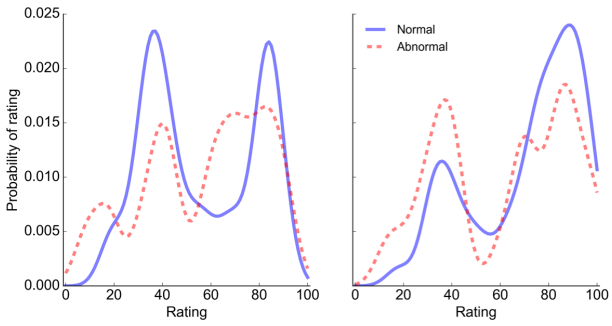

Figure 3: Examples of ratings given by two observers.

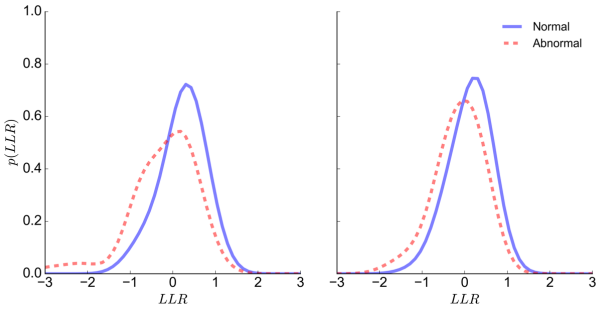

Figure 4: Distribution of log likelihood ratios (LLRs) computed from the data shown in Figure 3.

ROC curves and AUCs were computed separately for the real and synthesized conditions (see Figure 5). When the abnormal tissue did not contain a visible lesion, the performance given by the AUCs was similar across the real ($M = 0.64, SE = 0.01$) and the synthesized ($M = 0.65, SE = 0.01$) conditions, $t(10) = 0.40, p = 0.70$. However, when the abnormal tissue contained a visible lesion, the performance was better in the real condition ($M = 0.83, SE = 0.02$) than in the synthesized condition ($M = 0.71, SE = 0.02$), $t(7) = 5.12, p = 0.001$. In particular, performance increased for the real tissue sections.

To evaluate how the performance of each individual observer differs from chance levels, we characterized a figure of merit. For each observer, we created 100 Bootstrapped samples from the empirical data, derived a ROC for each sample, and computed the AUC. The 95th percentile of this distribu-

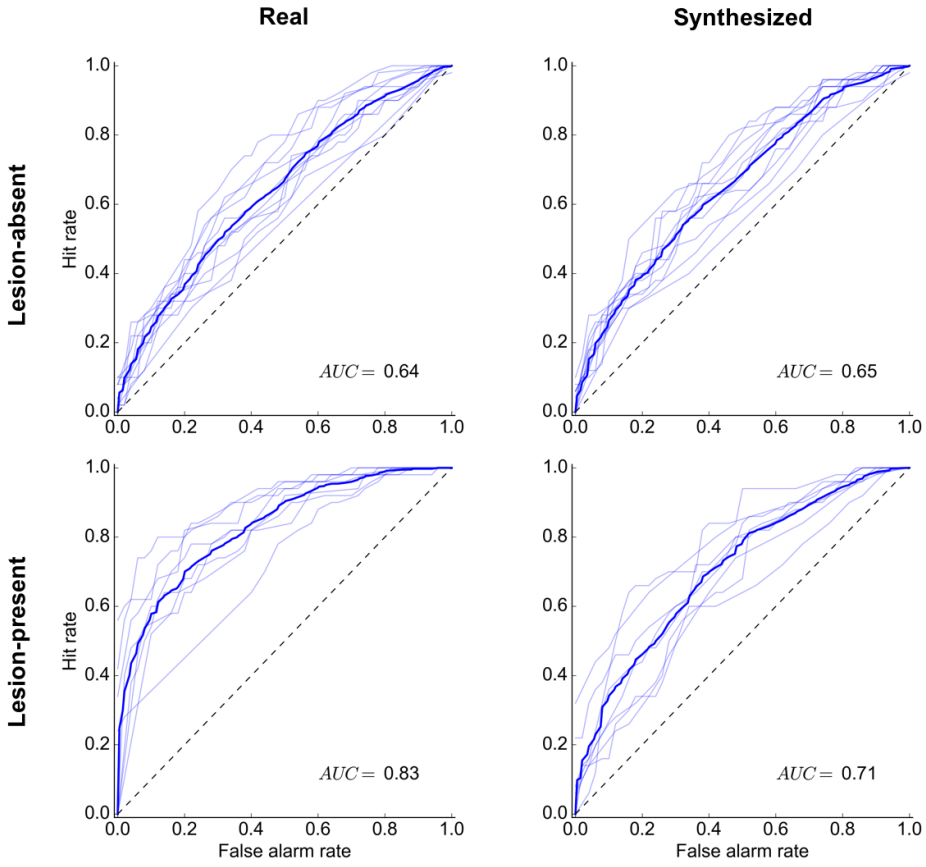

Figure 5: ROC curves in the real condition (left panel) and in the synthesized condition (right panel), when the abnormal tissue had no visible lesion (top panel) and when the abnormal tissue had a visible lesion (bottom panel). Thin lines represent the performance of individual observers. The thick line represents the average performance. The average AUCs are presented at the bottom right of each plot.

tion were set as the critical value. Figure 6 shows the resulting ROC curves for two individual observers.

When the abnormal tissue did not contain a visible lesion, two out of eleven observers performed above chance levels in the real condition while none of the eleven observers performed above chance levels in the synthesized condition. However, when a visible lesion was present in the abnormal tissue sections, all eight observers performed above chance levels in the real condition while two out of eight observers performed above chance levels in the synthesized condition.

Finally, to test whether radiologists gave similar ratings to the real and synthesized tissue sections, we compared the abnormality ratings. Regardless of the presence of a lesion in the tissue, we confirmed that the ratings were similar across the real and synthesized conditions (all $r > 0.27$, all $p < 0.001$).

Breast density analysis

Prior to this experiment, two radiologists gave clinical density ratings to the real tissue without any time limitations. To investigate whether the density of real tissue sections is perceived similarly in such short presentation, we compared the average clinical density ratings to the average perceived density ratings in our experiment. Positive correlations suggested that radiologists perceived the breast density similarly regardless of the duration of presentation (all $r > 0.59$, all $p < 0.001$).

Next, we investigated whether we were able to represent the perceived breast density of the real tissue sections in our synthesis. Comparison of the estimated density ratings across the real and synthesized conditions revealed a strong positive correlation ($r = 0.66, p < 0.001$), suggesting that our synthesis was successful in replicating the breast density.

Discussion

The purpose of the current study was to characterize the global signals of abnormality in breast tissue as texture descriptions (i.e., a set of stationary spatial statistics) and to determine whether radiologists rely on such texture descriptions when interpreting breast tissue. Our findings suggest that ra-

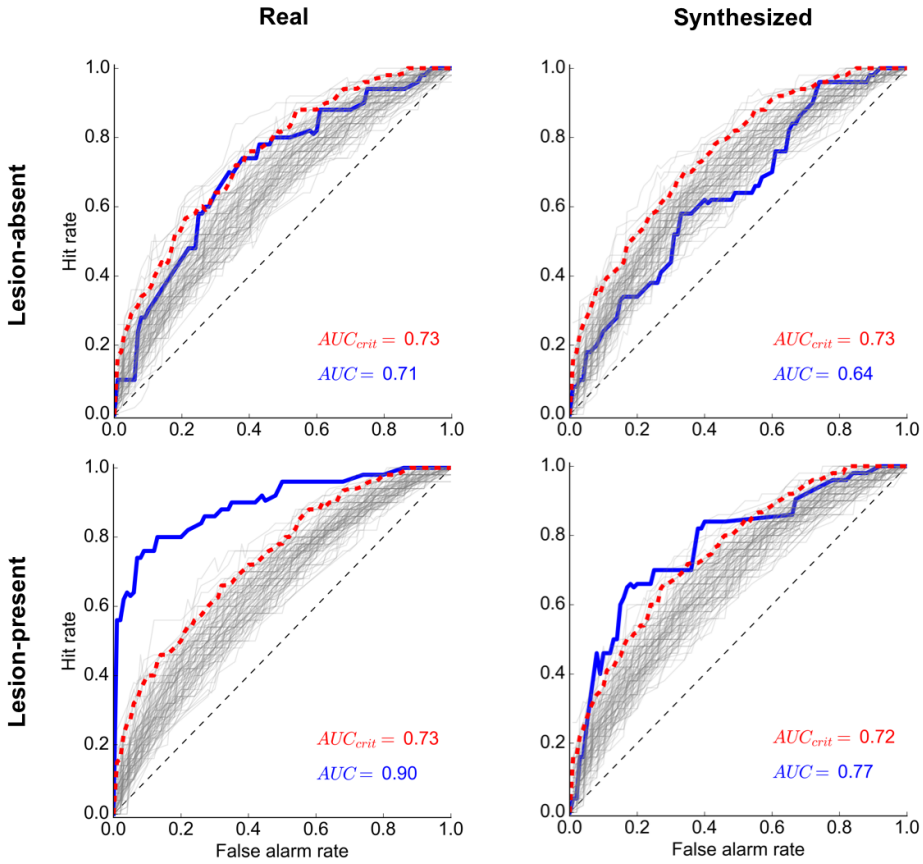

Figure 6: ROC curves for two of the observers. The top panel shows data from an observer when the abnormal tissue had no visible lesion, in the real condition (left panel) and in the synthesized condition (right panel). The bottom panel shows data from another observer when the abnormal tissue had a visible lesion, in the real condition (left panel) and in the synthesized condition (right panel). ROC curves for the Bootstrapped samples ($N = 100$) are given by the gray lines. The 95% of this distribution is given by the dashed red line. The empirical ROC curve is given by the blue line. The AUCs for the empirical (blue) and simulated (red) data are presented at the bottom left of each plot.

diologists can use texture descriptions as global signals of abnormality in diagnostic tasks.

When the abnormal tissue had no visible lesion, radiologists seemed to rely on texture descriptions; performance was similar across real and synthesized sections. However, when the abnormal tissue had a visible lesion, radiologists seemed to use additional mechanisms beyond the texture description. In particular, the existence of a lesion increased the performance only for the real sections. These findings confirm that radiologists can use texture descriptions as global signals of abnormality in interpretation of breast tissue.

Breast density judgments confirmed that the synthesized tissue represented the real tissue in terms of the breast density. Interestingly, the brief exposure time did not seem to influence the perception of breast density. Radiologists interpreted the breast density similarly with and without time limitations.

Using a similar paradigm in a series of experiments, Evans et al. (2016) showed that radiologists were able to discriminate between normal and abnormal breast tissue at a glance. To interpret this ability, they suggested that radiologists might be using some "global signal" of abnormality. In this study, we characterized this global signal as texture descriptions (i.e., a set of stationary spatial statistics) and confirmed that radiologists rely on such texture descriptions when interpreting abnormal breast tissue without a visible lesion.

In future work, we will examine the particular features of these texture descriptions that give rise to abnormality judgments. After determining the significant features, we will generate synthesized images using these particular features. Then, we will test these images in a similar paradigm where radiologists are asked to give abnormality judgments.

Overall, these findings contribute to the existing literature by suggesting texture statistics as global signals of abnormality in the interpretation of mammograms. There are several implications for improving detection of breast cancer. First,

relevant texture features can be used to synthesize "normal" or "abnormal" images representing breast tissue. These synthesized images can be used to train medical students or used as learning aids. Second, the features of texture descriptions can be used to train image classifiers to detect abnormality in the breast tissue, which ultimately could aid radiologists in the diagnostic process. Additionally, the synthesized images can be used in medical image perception research by allowing more control over the tissue samples, and by attenuating the limitations based on the small number of real cases.

Acknowledgments

This work was supported by NSF Grant BCS-1456822.

References

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Carmody, D. P., Nodine, C. F., & Kundel, H. L. (1980). Global and Segmented Search for Lung Nodules of Different Edge Gradients. *Investigative Radiology*, *15*(3), 224–233.

Drew, T., Evans, K., Võ, M. L. H., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images? *Radiographics*, *33*(1), 263–275. doi: 10.1148/rg.331125023

Evans, K. K., Georgian-Smith, D., Tambouret, R., Birdwell, R. L., & Wolfe, J. M. (2013). The gist of the abnormal: above-chance medical decision making in the blink of an eye. *Psychonomic bulletin & review*, *20*(6), 1170–5. doi: 10.3758/s13423-013-0459-3

Evans, K. K., Haygood, T. M., Cooper, J., Culpan, A.-M., & Wolfe, J. M. (2016). A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. *Proceedings of the National Academy of Sciences*, *113*(37), 10292–10297. doi: 10.1073/pnas.1606187113

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of vision*, *7*(1), 10. doi: 10.1167/7.1.10

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*(3), 316–355. doi: 10.1037/0096-3445.108.3.316

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176. doi: 10.1016/j.cogpsych.2008.06.001

Kundel, H. L., & Nodine, C. F. (1975). Interpreting Chest Radiographs without Visual Search. *Radiology*, *116*(3), 527–532. doi: 10.1148/116.3.527

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology*, *242*(2), 396–402. doi: 10.1148/radiol.2422051997

Kundel, H. L., Nodine, C. F., Thickman, D., & Toto, L. (1987). Searching for lung nodules. A comparison of human performance with random and systematic scanning models. *Investigative radiology*, *22*(5), 417–22.

Mugglestone, M., Gale, A., Cowley, H., & Wilson, A. (1995, apr). Diagnostic performance on briefly presented mammographic images. In H. L. Kundel (Ed.), *Proceedings of spie* (Vol. 2436, p. 106). International Society for Optics and Photonics. doi: 10.1117/12.206840

Nodine, C. F., Kundel, H. L., Lauver, S. C., & Toto, L. C. (1996). The Nature of Expertise in Searching Mammograms for Breast Masses. *In Medical Imaging*, *2712*, 89–94.

Oestmann, J. W., Greene, R., Kushner, D. C., Bourgouin, P. M., Linetsky, L., & Llewellyn, H. J. (1988). Lung Lesions - Correlation between Viewing Time and Detection. *Radiology*, *166*(2), 451–453. doi: 10.1148/radiology.166.2.3336720

Portilla, J., & Simoncelli, E. P. (2000). Aparametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–71.

Potter, M. C. (1976). Short-Term Conceptual Memory for Pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 509–522. doi: 10.1037/0278-7393.2.5.509

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Experimental Psychology*, *81*(1), 10–15. doi: 10.1037/h0027470

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, *5*(4), 195–200.

Swensson, R. G. (1980). A 2-Stage Detection Model Applied to Skilled Visual-Search by Radiologists. *Perception & Psychophysics*, *27*(1), 11–16. doi: 10.3758/BF03199899

Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77–84. doi: 10.1016/j.tics.2010.12.001