Learning and Confidence in 2D and 3D Medical Image Search

Maren Smith, Lauren Tavlan, Noelle Seim, Branden Song, Tal Sahar, Miguel A. Lago

Psychology and Brain Sciences, University of California Santa Barbara

Abstract

Humans search for specific targets in complex scenes to navigate the world. This ability to search is integral to survival in many ways, from its most basic role in hunting and gathering to its more advanced application to the detection of medical conditions in the field of radiology. According to previous research, the ability to search efficiently in a visual task can be learned over time. Despite sufficient evidence, in this paper, we recognize numerous findings that support the presence of greater learning and confidence curves in 3D versus 2D image search. The study of such learning patterns is important to the field of medicine as we hope to train radiologists to be as efficient, accurate, and confident as possible.

Introduction

Humans use visual search daily for survival and to navigate the world. Visual search is the act of searching for a specific target in a complex scene (i.e., finding food, avoiding predators, recognizing safe places, etc.). When performing visual search, people search the environment around them until they detect the target that they are looking for. In the field of radiology, medical image search is used to identify or rule out abnormalities that aid in the diagnosis of various medical conditions, such as the detection of malignant growths to confirm or deny the presence of cancer. Thus, medical image search proves to be an invaluable tool that enables medical professionals to save lives. However, searching in this type of imaging modality requires heavy training before one is able to efficiently search and recognize what the target looks like.

According to previous research, the ability to search effectively and efficiently in a visual task can be learned over time [1, 2]. These findings extend to radiologists' searches for cancer nodules in medical images. It has been shown that long-time radiologists tend to have lower recall rates than radiologists that have worked for less than three years [3], and newer radiologists demonstrate improvements in their false-positive rates over time [4]. As training and experience are gained, confidence levels in performance also improve [5].

When applying learning curves, efficiency rates, and confidence levels to radiology images, it is important to also consider the differences within the types of images analyzed. The search strategies employed in both two-dimensional full field digital mammography (FFDM) and three-dimensional digital breast tomosynthesis (DBT) are fundamentally different [6]¬¬¬– although they are both used to identify possible breast cancer causing nodules. Despite this difference, performance in DBT and FFDM has indicated no significant differences [7]. However, more evidence reveals that searching in three-dimensional images leads to an increase in accuracy within the true positive rate in DBT compared to FFDM [8, 9]. Application of these differences to learning curves have not been extensively studied, but a recent study showed that within-radiologist learning can occur in DBT [10], displaying potential for the demonstration of learning curves in medical images. Given that these image types are used to identify breast cancer agents, understanding the disparities in their learning curves, efficiency rates, and confidence levels is an area of interest for our research.

The goal of this paper is to analyze the impact of learning on correct cancer nodule identifications in both two-dimensional and three-dimensional medical images and its subsequent effects on efficiency and confidence levels. We hypothesize that, as learning occurs, the percent of correct identifications will increase, along with the efficiency of each search and the participants' confidence. We believe these findings will occur more notably in three-dimensional images compared to two-dimensional images.

Materials and Methods

Participants

Participants in this study consisted of 11 students from the University of California, Santa Barbara. There were 8 female and 3 male participants between the ages of 20 and 23. All participants received school credit upon completion of the experiment and academic quarter and were provided informed consent. Participants had verified normal or corrected-to-normal vision.

Stimuli

Two different types of signals were utilized in the experiment: masses and microcalcifications. The masses consisted of a 2D/3D Gaussian blob with a standard deviation of 10 pixels/voxels (roughly 2.6 mm). The microcalcifications consisted of a sharp, defined sphere with a 6 pixel/voxel (roughly 1.56 mm) diameter. Additionally, the microcalcifications had constant contrast equal to the maximum contrast of the mass (Figure 1). In the 2D/3D images, the signals were present in 50% of the trials in random locations on a generated noise field, while the other 50% of the trials did not have a stimulus present. Of the trials with a stimulus present, 50% of the trials asked the participant to look for a mass, and 50% of the trials asked the participant to look for a microcalcification. The background of the stimuli was generated with a 2D/3D correlated Gaussian noise field that resembled the noise present in medical breast mammograms [11].

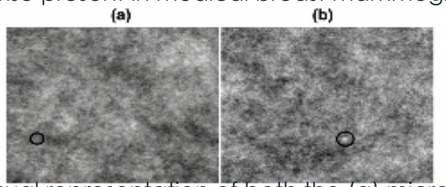

Figure 1. Visual representation of both the (a) microcalcification and (b) mass signals that participants were prompted to search for in each trial.

Design and Procedure

This between-subjects experiment consisted of 800 trials per participant (400 for 2D, 400 for 3D). The manipulated variable was the type of stimulus present or non-present in each trial: mass-present, mass-non-present, microcalcification-present, and microcalcification-non-present.

All the participants individually entered the testing room, where they were directed to a table that included a chinrest, two monitors, and a computer mouse. The chin rest was adjusted to a comfortable setting, allowing the eye tracker to track eye movements with greater ease. Two monitors were used for the experiment and placed side by side. The monitor on the left controlled the eye tracker, as well as recording and saving eye movements throughout the study. The monitor on the right was a medical-grade monitor calibrated linearly (Barco MDRC 1119) that presented the different stimuli in the trials.

Participants placed their chin on the chin rest and verified that the eye tracker was tracking only the right eye. Next, participants opened the program on MATLAB, calibrated the eye tracker, and validated it, ensuring accurate and precise tracking. The light was turned off, and participants were instructed on how to use the mouse and its jog wheel to navigate the trials. After a short round of practice trials, and when participants demonstrated understanding, they pressed the escape key to start the experiment.

Participants alternated between 2-dimensional and 3-dimensional search modalities. At the beginning of each trial, they maintained gaze at the center of the screen through forced fixation and initiated the start of the trial by pressing the spacebar. The participant then scanned the image to determine whether a mass or microcalcification was present. Each trial was target-specific, meaning that participants knew whether they were to search for a mass or microcalcification before the start of each trial. If the stimulus was present, participants pressed the spacebar and further denoted the confidence with which they believed the target to be present via a confidence rating of 5 to 8, 5 being the lowest confidence and 8 being the highest confidence. On the other hand, if the participant believed the target to be absent, they would rate their confidence that the stimulus was absent through a confidence rating of 1 to 4, 1 being the highest confidence, and 4 being the lowest confidence. They were provided feedback with the correct answer and, in the case of a present trial, the location of it. Figure 2 shows an example

of the timeline for one trial.

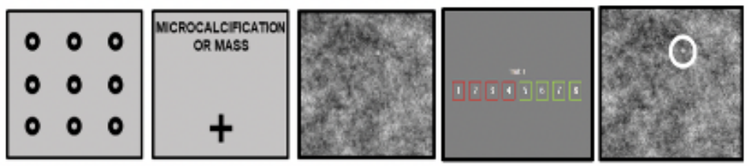

Figure 2. A single trial's visual timeline in the experiment, including calibration, primary fixation, noise field search, confidence rating, and revealed present target.

The 3-dimensional trials required the participants to scroll through a "stack" of images using the mouse's wheel to scroll through different slices of the volume while attempting to locate the specified mass or microcalcification. Participants followed the same procedure used in the 2-dimensional trials to terminate the trial and respond.

Figures of Merit

Proportion correct (PC) was calculated as follows:PC=(TP+TN)/N, where TP stands for the number of true positives, TN stands for the number of true negatives, and N represents the total amount of trials attempted. Trial time is the amount of time that it took the participant to confirm or deny the presence of either a microcalcification or a mass within a given trial. The trial time duration was measured, in seconds, from the start of the stimulus display until the participant ended the trial by hitting the spacebar. Efficiency was calculated as the ratio between PC and trial time: PC/Trial Time.

Results

Figure 3 depicts the proportion of correct trials (averaged across 11 participants) for each of the 400 trials attempted. In both 2D and 3D images, it appears that participants can more accurately identify or reject the presence of a microcalcification than a mass, regardless of trial number. While this accuracy disparity is present in both types of images, it is only significant ($p<0.05$) in 2D images, as depicted by the presence of dots in Figure 3a. In the 3D search, it appears that this accuracy disparity between stimuli lessens with an increasing trial number, as illustrated by the lack of dots in Figure 3b. More so, nearly every set of data points is significantly different from one another in the 2D images, while for the 3D images, not a single set of data points is significantly different from its partner. For both types of images, PC increases with the trial number. In the 2D search, it appears that PC for both microcalcifications and masses increase until about trial numbers 30 and 50, respectively, after which it begins to plateau. In 3D search, the PC for both microcalcifications and masses appear to increase significantly within the first ~50 trials, as also seen in 2D search. However, 3D PC seems to gradually increase with the trial number instead of plateauing like 2D PC.

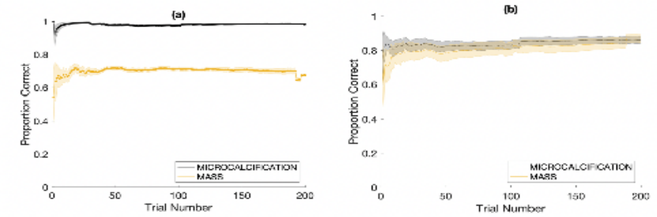

Figure 3. Proportion correct versus trial number for a (a) 2D and (b) 3D image. These curves represent the average of all 11 participants for both microcalcifications and masses. Shaded areas are standard errors of the mean. The dots in the plots show which set of data points are significantly different from each other ($p<0.05$).

With regards to accuracy (PC), for both 2D and 3D images, it appears that microcalcifications consistently present a higher proportion correct (PC) than masses do, regardless of trial number. This suggests that people are perpetually better at correctly identifying or rejecting microcalcifications over masses. In alignment with previous literature, this finding supports the notion that there are discrepancies between 2D and 3D search accuracy. The gradual increase in 3D PC over growing trial numbers indicates that participants become better at correctly identifying or rejecting both microcalcifications and masses in 3D as time goes on. In the case of 2D images, the PC for microcalcifications and masses plateaus around trial numbers 30 and 50, respectively, meaning that participants do not get any better or worse at correctly identifying or rejecting microcalcifications or masses in 2D after this point. These findings support our hypothesis of a greater 3D learning curve and reinforce the existence of modality-specific learning curves with regards to the medical image search. The abundance of "dots" on the 2D plot demonstrates that participants are consistently and significantly better at correctly identifying microcalcifications over masses in 2D. This is not the case in 3D, as this plot contains no dots. Furthermore, it is apparent that participants are better, and nearly perfect at correctly identifying microcalcifications in 2D rather than 3D, as indicated by their respective accuracy percentages of around 96% and 83%. The opposite appears to be true for masses, with relative 2D and 3D PC percentages of around 73% and 78%. These disparities are more examples of

how there are distinctive patterns of improvement dependent on search modality. These effects were not seen when absent only trials were analyzed.

Figure 4 shows the amount of time (averaged across 11 participants) that it took to confirm or deny the presence of either a microcalcification or a mass within a given absent trial. For both types of images, it appears that microcalcification trial time is consistently higher than that of the masses, regardless of trial number. This disparity is more significant in 3D search, especially in the final quarter of trials, as depicted by the abundance of dots in this section of the plot. Overall, trial time is greater in 3D images than in 2D images. In both types of search, trial time appears to decrease with a rising trial number.

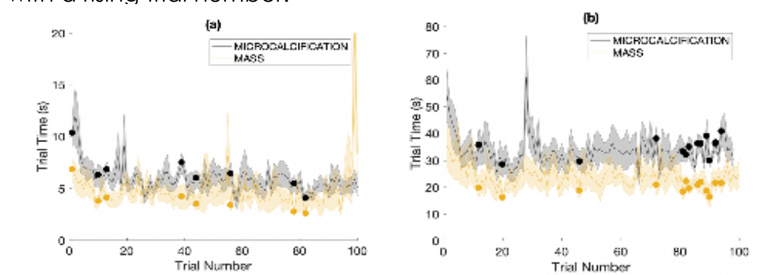

Figure 4. Trial time versus trial number for (a) 2D and (b) 3D search, using only absent trials. These curves represent the average of all 11 participants for both microcalcifications and masses. Shaded areas are standard errors of the mean. The dots in the plots show which set of data points are significantly different from each other ($p<0.05$).

In studying trial time, we see that the amount of time spent analyzing a 3D image for a microcalcification begins to separate significantly from the amount of time spent searching for a mass in the final quarter of trials¬¬– with a larger amount of time spent on trials in which the participant was prompted to search for a microcalcification. This finding illustrates how participants can more quickly recognize the presence or absence of a mass over a microcalcification during a 3D search. This disposition further supports our hypothesis of a larger 3D learning curve and illustrates the presence of stimuli discrepancies within the same image type. The increased amount of visual data that is processed in 3D compared with 2D images is thought to be the reason why participants spent much less time searching through 2D images than 3D images. For both image types, regardless of stimuli, trial time decreases slightly as the trial number increases, suggesting that participants get more confident and comfortable with identifying both microcalcifications and masses as time goes by. These effects were not seen when all trials were analyzed.

Figure 5 illustrates the efficiency (averaged across 11 participants) for each of the absent trials attempted by the participants. Overall, efficiency is higher for 2D search (~16%-18%) than 3D search (~3%-6%), regardless of stimulus type. In the 2D images, it appears that there is no efficiency discrepancy between microcalcifications and masses, as illustrated by the absence of dots in Figure 5a. However, in 3D search, it appears that mass efficiency is greater than microcalcification efficiency, as depicted by the increased number of dots in Figure 5b. This efficiency disparity appears to increase around trial number 67 as mass search efficiency begins to rise and separate significantly from microcalcification search efficiency. In both 2D and 3D search, efficiency increases with the trial number.

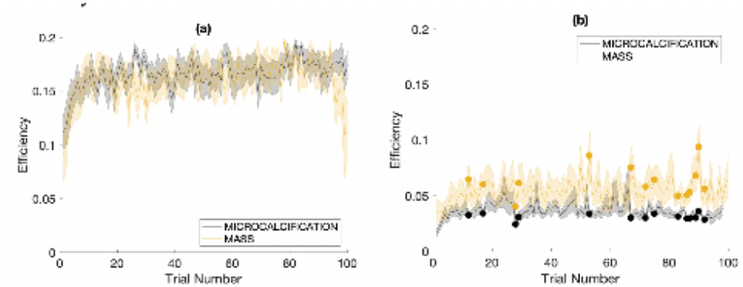

Figure 5. Efficiency versus trial number for (a) 2D and (b) 3D search, using only absent trials. These curves represent the average of all 11 participants for both microcalcifications and masses. Shaded areas are standard errors of the mean. The dots in the plots show which set of data points are significantly different from each other ($p<0.05$).

When we study efficiency, it appears that for 3D images, participants are almost always more efficient at accurately identifying masses over microcalcifications, regardless of the trial number. This is not observed in the 2D images and suggests that participants are more efficient at searching for masses over microcalcifications when they are searching a 3D image but not when they are searching a 2D image. This finding further supports our hypothesis of a larger 3D learning curve and emphasizes earlier results that these learning patterns can not only be image-specific but also unique to the stimulus found within a given image type. For both the 2D and 3D images, it appears that overall efficiency gradually increases and eventually plateaus as the trial number increases. This finding indicates that participants get more effi-

cient at correctly identifying both microcalcifications and masses as time goes on, suggesting the presence of a learning process in both search modalities¬¬. When comparing 2D and 3D images, it is evident that overall efficiency is much higher for 2D search– meaning that participants are much more efficient at correctly identifying both microcalcifications and masses in 2D rather than in 3D. This data aligns with previous literature in supporting the notion that there are mode-specific learning curves as well as discrepancies between 2D and 3D search accuracy. These effects are better observed when analyzing absent only trials.

Figure 6 illustrates the confidence with which participants identified or rejected the presence of a signal as a function of trial number. The average of all 11 participants for each of the 400 trials attempted are analyzed in this figure. For the 2D microcalcification search, it appears that the 8 rating is consistently high throughout all 200 trials and that the amount of 2 ratings increases with the trial number. Additionally, it appears that the number of 3 ratings decreases as the trial number rises. For 2D mass search, it appears that the 1 and 4 ratings increase with trial number while the 6 rating decreases with the trial number. For the 3D microcalcification search, it appears that the 3 and 8 ratings increase with trial number while the 4 and 5 ratings decrease with increasing trial number. For 3D mass search, it appears that the 1, 4, and 7 ratings increase with trial number while the 5 and 6 ratings decrease with a rising trial number.

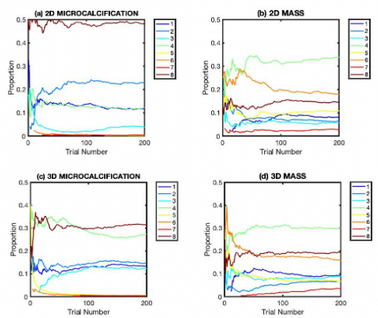

Figure 6. The proportion of a given confidence rating versus trial number for a 2D image containing either a (a) microcalcification or a (b) mass and a 3D image containing either a (c) microcalcification or a (d) mass. These curves represent the average of all 11 participants for both microcalcifications and masses. The different colors represent the various confidence ratings (1-8).

The final analysis focuses on confidence ratings for both stimuli identification and stimuli rejection as a function of trial number (Figure 6). In the case of 2D microcalcification search, the results suggest that participants are consistently confident in their ability to correctly identify a microcalcification in a 2D image. Additionally, it appears that with a rising trial number, participants become increasingly more confident in their ability to correctly reject the presence of a microcalcification. It is important to denote that, although the data shows increasing signs of improvement, the number of 1 ratings did not increase significantly with the trial number. This means that participants become more confident in their ability to accurately identify rather than reject the presence of a microcalcification in 2D as time goes on. When analyzing the confidence ratings for 2D images containing masses, it appears that participant's confidence in their ability to correctly reject the presence of a mass increases with time. This finding is opposite of that seen for the 2D microcalcification search, suggesting that as participants spend more time searching in 2D, they become more confident in their ability to correctly identify a microcalcification and reject a mass. Cohesive with the findings concerning 2D PC and efficiency, it is apparent that for 2D search, as participant confidence increases, so does their accuracy (PC) and efficiency.

When analyzing 3D images containing microcalcifications, it appears that as participants spend more time in 3D search, they become more confident in their ability to correctly identify the presence of a microcalcification. When analyzing 3D images containing masses, it is apparent that with increasing trial numbers, participants become more confident in their ability to both correctly identify and reject the presence of a mass. However, as mentioned above, it is important to note that although the data for 3D mass search shows increasing signs of improvement, with more time spent in 3D search, participants become more confident in their ability to accurately reject rather than confirm the presence of a mass. Cohesive with the findings from 3D PC and efficiency, it is apparent that for 3D search, as participant confidence increases, so does their accuracy (PC) and efficiency.

Overall, this study of confidence ratings suggests that regardless of image type (2D or 3D), participants become more confident in their ability to correctly identify the presence of a microcalcification and reject the presence of a mass as time goes on. Additionally, it appears that the noise present at the beginning section of trials dissipates with increasing trial number, hence the smoother lines. This alludes to the notion that as trial numbers increase, participant's ability to habituate to the search process and the use of the confidence scale increases as well. We believe this

to be yet another measure of learning that participants express. With respect to all of these results, it is clear that as participants learn (increase their proportion correct and efficiency), their confidence increases as well. These results not only support our hypothesis but also complement our previous findings in upholding the notion that there are discrepancies between 2D and 3D learning curves. Through this study of confidence ratings, we were able to illustrate that these differences go beyond accuracy, efficiency, or trial time and include other aspects of image search such as confidence.

Conclusion

While the evidence is not sufficient enough to fully reject or accept our hypothesis, there are some findings to support the notion of a greater 3D learning and confidence curve. When looking at efficiency, we observe that participants learn how to search for masses in 3D more so than they do in 2D. Furthermore, when comparing accuracy, we observe that the learning process in 2D image search is reduced to ~50 trials while the 3D learning process continues throughout all 200 trials. Additionally, when comparing confidence ratings, it appears that with time participants become more confident with 3D image search rather than 2D image search. This demonstration of a larger 3D learning curve supports previous literature in upholding the notion that there are accuracy differences between search modalities as well as modality-specific learning curves. Furthermore, there is evidence that efficiency and confidence, for both types of image searches, appear to be affected by the type of cancer nodule being searched for. However, our findings make it clear that 2D search efficiency exceeds that of 3D search, regardless of stimulus type.

While this study was limited by small sample size, a sample of untrained medical professionals, and inauthentic noise generated 2D and 3D mammograms, this experiment illustrated that there are disparities in efficiency and confidence between 2D and 3D medical image search and that more research is needed to fully understand the extent and possible cause of these differences. This incorporation of confidence ratings to the study of modality-specific learning curves leads us to wonder if there is a definitive relationship between the two. Exploring the details of this relationship with regards to what is already known about both search modalities is a possible area of future research. Understanding the logistics and differences between the ways in which people learn to search both 2D and 3D medical images holds true importance as it may shape the way medical professionals in the field are trained. With a deeper knowledge of the mechanisms that make 2D and 3D medical image search different from one another, we hope to train future radiologists to be the most efficient, accurate, and confident as possible so that they can better diagnose, treat and care for their patients.

References

[1] Sireteanu, R., & Rettenbach, R. (1995). Perceptual learning in visual search: Fast, enduring,

but non-specific. Vision research, 35(14), 2037-2043.

[2] Ericson, J. M., Kravitz, D. J., & Mitroff, S. R. (2017). Visual search: you are who you are (+

a learning curve). Perception, 46(12), 1434-1441.

[3] Miglioretti, D. L., Gard, C. C., Carney, P. A., Onega, T. L., Buist, D. S., Sickles, E. A., ... &

Elmore, J. G. (2009). When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology, 253(3), 632-640.

[4] Barlow, W. E., Chi, C., Carney, P. A., Taplin, S. H., D'Orsi, C., Cutter, G., ... & Elmore, J. G.

(2004). Accuracy of screening mammography interpretation by characteristics of radiologists. Journal of the National Cancer Institute, 96(24), 1840-1850.

[5] Clanton, J., Gardner, A., Cheung, M., Mellert, L., Evancho-Chapman, M., & George, R. L.

(2014). The relationship between confidence and competence in the development of

surgical skills. Journal of surgical education, 71(3), 405-412.

[6] Aizenman, A., Drew, T., Ehinger, K. A., Georgian-Smith, D., & Wolfe, J. M. (2017).

Comparing search patterns in digital breast tomosynthesis and full-field digital

mammography: an eye tracking study. Journal of Medical Imaging, 4(4), 045501.

[7] Gennaro, G., Toledano, A., Di Maggio, C., Baldan, E., Bezzon, E., La Grassa, M., ... &

Muzzio, P. C. (2010). Digital breast tomosynthesis versus digital mammography: a

About the Author

Maren Smith is a fourth-year Biopsychology major who is graduating this spring. Outside of the classroom, Maren is a medical scribe at Cottage Hospital and plays on the Women's Club Soccer team. After UCSB, she plans to pursue a career in medicine.