https://doi.org/10.2352/ISSN.2470-1173.2019.12.HVEI-225

Do different radiologists perceive medical images the same way? Some insights from Representational Similarity Analysis

Jay Hegdé; Augusta University; Augusta, Georgia/USA; and Evgeniy Bart; Palo Alto Research Center, California/USA

Abstract

Characterizing what experts perceive in medical images is a difficult problem, both because doing so requires somehow characterizing the internal mental representations of the observer, and because the underlying diagnostic information tends to be abstract and not readily describable in terms of well-defined image features. Representational Similarity Analysis (RSA) is a method originally developed in mathematical psychology that provides a theoretically sound and quantitative framework for measuring the mental representations of visual images in human observers. Here we used RSA to measure the extent to which the same underlying set of mammograms elicit similar mental representations in different practicing radiologists (N = 26). We found that the internal representations were statistically indistinguishable across different radiologists ($p > 0.05$). Moreover, the mental representations significantly parallel the diagnostic information in the images ($p < 0.05$ for each subject), indicating that various radiologists perceived the same set of diagnostic information in the underlying images. Together, these results indicate that medical images elicit similar mental representations in different radiologists.

Introduction

It is a truism of public health policy that the health outcomes for a given patient should be roughly similar regardless of who the patient's clinical providers are. Indeed, reducing the variability across health providers has been widely recognized as a crucial part of improving outcomes for patients [1,2]. However, as an empirical matter, health outcomes do vary considerably [3]. Therefore, understanding how and where the variability arises is crucial to reducing it.

In medical specialties in which medical image perception plays a key role, such as radiology and pathology, an obvious starting point is to measure the extent to which clinicians vary in the way they perceive medical images. After all, the complexities and ambiguities of radiological images are known to be key contributing factors to the intra- and inter-observer variability, and to medical errors, in these fields [4-12].

The problem of characterizing high-level perception of images

It is intuitively obvious that diagnostic image patterns in medical images tend to be subtle and abstract – if they were not, it would not take highly trained clinical experts (or, to duly ingratiate the present readership, machines) to carry out the diagnostic task; anyone would be able to do it. It stands to reason that the internal (*i.e.*, mental) percepts generated by these images in the medical expert must also be subtle and abstract. When the underlying percept is so abstract, how does one go about quantitatively characterizing it and comparing the relevant percepts across multiple subjects to boot?

At first glance, this problem may seem intractable. To appreciate the difficulty, consider the four images in Figure 1. The images are all quite different from each other with respect to their low-level (or in Shepard's terminology, 'first-order') image characteristics, such as color, luminance, local contrast, etc., so that the percepts generated by the first-order characteristics will vary greatly depending on the image, even within a single subject. But what do all four images have in common? A close scrutiny of the image will make it evident that the answer is that the images all contain representations of the number two (see legend for details).

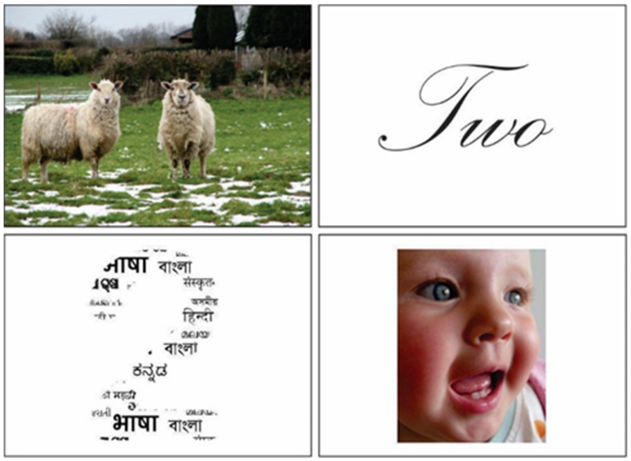


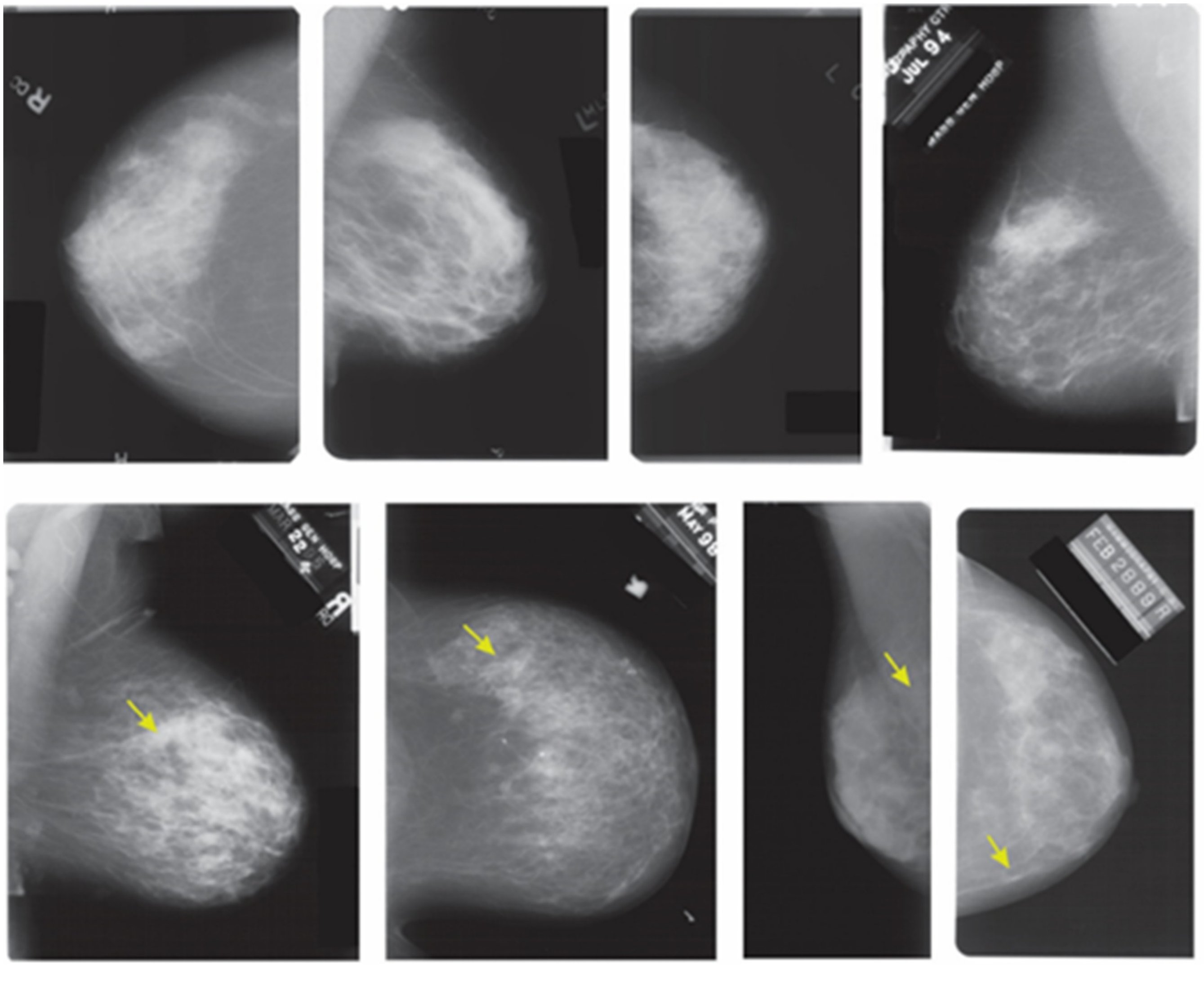
***Figure 1.** The problem of measuring mental representations of abstract visual features. It is readily apparent that the four panels in this figure share very little of the low-level visual characteristics (i.e., first-order image statistics), such as local orientation, contrast, color, etc. Therefore, they are said to lack first-order isomorphisms (i.e., similarities) [13]. But at a more abstract, 'second-order' level, all four images generate an internal representation of the number two, such as the two lambs, or the two eyes or the two little baby teeth of the baby, etc. Thus, the images share a more functional, second-order isomorphism [13,14], regardless of functionally irrelevant first-order image variations (or polymorphisms) [14]. In this sense, the representations generated by this figure are directly analogous to the radiologist's internal representations of diagnostic features, where the underlying images tend to be physically quite different even when the underlying diagnosis is the same. Characterizing such abstract internal representations may appear to be an intractable problem. However, decades of research on representational similarity has established theoretically sound, effective methods for measuring such second-order, functional representations, and factoring out the confounding contributions of first-order representations [13-15,17,18,20-24]. The work described here leverages this approach to help characterize the representation of diagnostic features in mammograms. Figure adapted from Fig. 1 of Shepard et al [13]. Images courtesy of Wikimedia Commons.*

Shepard's brilliant intuition was that to the extent that one is able to perceive these high-order similarities, the underlying higher-level mental representations of the four images must be similar, or 'isomophic' [13-15]. And, to the extent different people perceive the same high-level similarities, the corresponding mental representations must also be isomorphic across subjects. Several decades of subsequent work has validated the RSA framework and extended it to a variety of areas, including correlating perceptions with underlying brain activity [16]. RSA has been used to psychophysically characterize the internal representation of numbers in human subjects [13], and the representation of object shapes in human subjects [17] and in monkeys [18]. It has been used for similar purposes in a large number of human neuroimaging studies (for a review, see ref. [16]). Therefore, comparing mental representations of medical images across different radiologists represents a principled, but novel, application of RSA to the study of medical image perception.

Using RSA to medical image perception: A brief outline of our RSA methodology

Here we describe the application of RSA to the study of medical image perception using mammograms as an illustrative case (Figs. 2 and 3; see legends for details). All subjects were adult volunteers who provided informed consent. All protocols used in this study were reviewed and approved in advance by the Institutional Review Board (IRB) of Augusta University (AU). Subjects were tested either at AU or at the Perception Laboratory during one of the annual meetings of the Radiological Society of North America (RSNA). Briefly, subjects viewed a given randomly drawn pair of mammograms, each of which had a 50% probability of containing a cancer, and rated their dissimilarity. The reason for using dissimilarity (as opposed to similarity) is that by this measure, percepts that are mutually similar, *i.e*., close to each other in the perceptual space, will have a smaller Euclidean distance between them when plotted [16]. By repeating the pairwise comparison for each pair *i,j* of n mammograms, we constructed a diagonally symmetric n x n dissimilarity matrix (RDM), where cells *i,j* and *j,i* contain the dissimilarity rating of the corresponding pair of stimuli.

Figure 2. *Exemplar mammograms. These arbitrarily chosen 2-D mammograms help illustrate the diagnostic complexity and variability of mammograms. Each of the four mammograms in the top row is healthy, and each of the mammograms in the bottom row contains a cancer. A few features of this image set are worth noting. First, note the variability among the images. As an empirical matter, two given mammograms are never identical, even if they are of the same breast taken during the same visit [19,25]. Second, to the untrained eye the two sets of mammograms seem indistinguishable. Third, the cancerous regions (arrows) are not necessarily the most salient regions of the image, and vice versa. Finally, the area of the cancerous region generally accounts for a tiny fraction (typically 1-2%) of the overall area of the mammogram, so that, even in a cancerous breast, most of the rest of the tissue is actually healthy. For all these reasons, analyses of low-level image characteristics typically fail to accurately distinguish cancerous mammograms from healthy ones, and vice versa. For the same set of reasons, the two sets of mammograms appear perceptually indistinguishable to the untrained eye.*



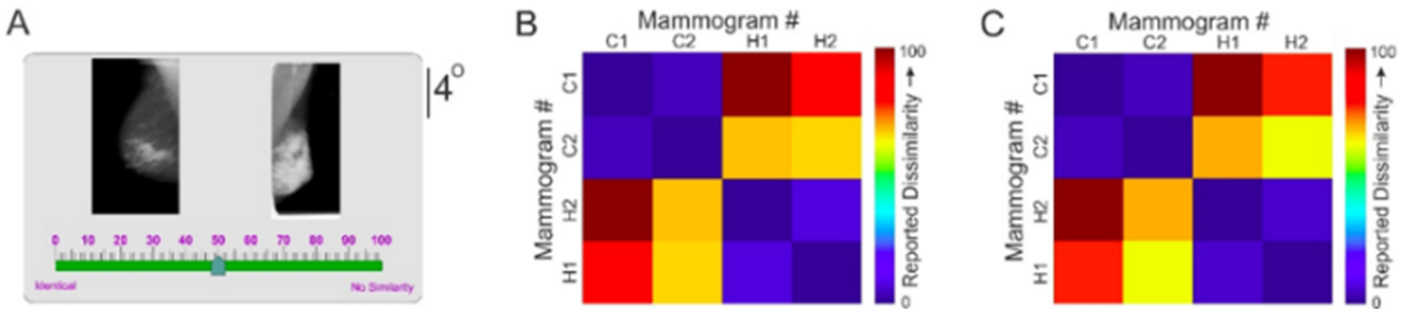


Figure 3. *Steps in the RSA methodology. (A) Trial paradigm. Each trial started when the subject fixated a central fixation spot (not shown) and indicated trial readiness by pressing a key. A randomly drawn pair of mammograms were presented on a high-resolution monitor for ad libitum viewing. Subjects were required to make a graded report of the perceived dissimilarity between the two mammograms using a slider (bottom), and press a separate key (not shown) to confirm the report. Figure not drawn to exact scale. For additional details, see refs. [26,27]. No systematic relationship was evident between the reaction time (i.e., the time the subjects took before responding) and the reported dissimilarity (data not shown). 320 radiologically vetted mammograms (half of which were cancerous, i.e., contained a single cancer, the remaining half being healthy) were used. Each subject was tested with eight healthy and cancerous mammograms each, and the mammograms were systematically counter-rotated across subjects. For each subject, reported dissimilarity for all possible pairs of the16 mammograms were used to construct a 16x16 representational dissimilarity matrix (RDM). For clarity, a 4x4 subset of this larger RDM is shown for two different subjects in panels (B) and (C). C1 and C2 denote two different cancerous mammograms, and H1 and H2 denote two different healthy mammograms.*

4x4 RDMs for a representative subset of 4 mammograms (a pair each of healthy and cancerous mammograms) are shown for two different subjects in Fig. 3B and C (see legend for details; also see also see Figs. 2-3 of ref. [16]). The similarity between the RDMs, measured using a Congruence Coefficient *C*, is a numeric measure of the extent to which the internal representations of the two subjects were similar.

Mammography experts perceive mammograms highly similarly

Using this approach, we measured the internal representations of 26 practicing radiologists who specialized in mammography and had at least 12 years of mammography experience (Fig. 3; see legend and ref. for methodological details). For each possible pair of these 26 mammography experts, we determined pairwise congruence coefficients *C*. The higher congruence coefficient values, the greater the similarity between the underlying mental representations [13,16].

For cancerous mammograms, the similarity of reported percepts was highly significant (leftmost bar in Fig. 4). The similarity was also high for healthy breasts, but the variance was larger (middle bar). Indeed, the variance of the reported percepts was significantly higher for healthy breasts than for cancerous breasts (*F* tests, $p < 0.05$; data not shown). While the present study did not address the reason for this, one plausible scenario is that for cancerous breasts, the experts focus on the region of interest (ROI) containing cancer to determine their dissimilarity ratings. Such ROIs typically account for no more than 1-2% of the overall image area (data not shown). Since the cancer-containing ROIs are, by definition, diagnostically highly similar, they tend to elicit correspondingly similar ratings. By contrast, when the breasts are healthy, there is no spatially restricted ROI, and the dissimilarity ratings will be based on the entire breast, so that variability in the images is reflected in the dissimilarity ratings.

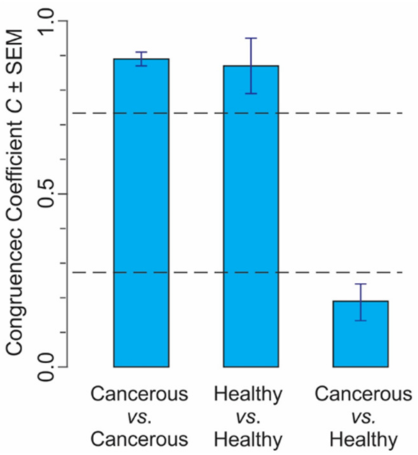


Figure 4. *Similarity of reported mammogram percepts for experienced radiologists. All subjects (N = 26) were mammography specialists with at least 12 years of experience in mammography. For each of the three categories of mammogram comparison (cancerous vs. cancerous, healthy vs. healthy, and healthy vs. cancerous), similarity between the reported percepts among subjects was measured by calculating the pairwise Congruence Coefficient C 18 for all possible pairs of the subjects. The average C ± SEM for all 26 subjects for each mammogram category is shown in this figure. The dotted lines denote the 95% confidence intervals for this dataset, as determined by randomization, corrected for multiple comparisons [18].*

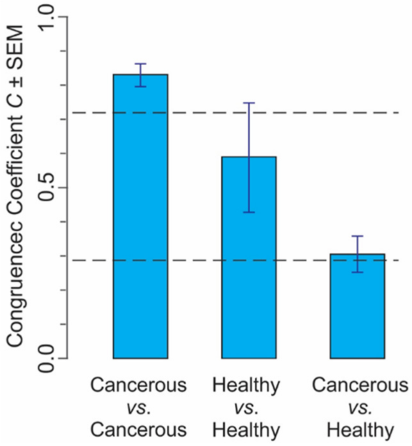


Figure 5. *Similarity of reported mammogram percepts for radiology residents (N = 22). Average C ± SEM for the residents were determined as plotted in this figure using the same procedure as was used for experienced radiologists (see legend to Fig. 4).*

Expert radiologists perceived cancerous mammograms significantly differently from healthy mammograms (rightmost bar). This suggests that the internal representations of expert radiologists are driven primarily by the diagnostic information in the underlying images, so that the mental representation within each diagnostic category (*i.e.*, cancerous or healthy) is highly similar, and significantly different across the categories.

Reported percepts of radiology residents are highly variable

We repeated the above experiment using 22 radiology residents, only three of whom specialized in mammography. These subjects had 0.5 to 2.5 years of experience as a radiology resident, depending on the resident. The reported percepts were significantly similar across subjects for none of the aforementioned three categories (Fig. 5). The congruence with image information was also insignificant for each individual subject ($p > 0.05$ for each subject; data not shown). Interestingly, the similarity between the image information and the corresponding mental representation was significantly correlated with the subject's expertise ($r = 0.63$; $df = 18$; $p < 0.05$).

The present study did not address the reasons for the lack of significant representational similarity among radiology residents. A plausible explanation is that the residents are still in the process of learning the abstract image patterns diagnostic of breast cancer, and are therefore not able to fully differentiate cancerous images from healthy images, ignore non-diagnostic variability, or both. A second, non-exclusive scenario is that the variability of the internal representations reflect the underlying variability of this subject sample. After all, the residents specialized in a broad variety of radiological sub-specialties other than mammography. However, one line of evidence that lends support to the notion that the level of expertise was related to the observed convergence of mental representations among experts is that, across all radiologists including aforementioned experts and residents, the level of congruence of mental representation was significantly correlated with the subject's expertise ($r = 0.57$; $df = 46$; $p < 0.05$).

Conclusions

Our results demonstrate that RSA is an effective method for quantitatively comparing the internal representation of medical images. They also demonstrate that different mammography specialists perceive mammograms similarly, and that the similarity of perception depends on the level of expertise.

Whether and to what extent our results generalize to other subspecialties of radiology or other specialties involving medical image perception remains to be studied. Our preliminary data (not shown) indicate that when the image itself plays a comparatively smaller role in cancer diagnosis, e.g., inflammatory breast cancer, the similarity of mental representations tends to be smaller and more variable. Our results (not shown) also indicate the similarity of internal representation is proportional to clarity of the diagnostic information in the mammogram. For instance, the representations were more similar when the subjects were shown radiologically vetted mammograms that were classified as "Highly suggestive of malignancy" (using the standard BI-RADS classification scheme [19]), and less similar when the subjects were shown mammograms that belonged to the less definitive diagnostic category of "Suspicious". Most importantly, our study provides the proof of the principle that RSA provides a principled, quantitative and theoretically sound method of studying medical image perception.

Acknowledgements

We thank Ms. Fallon Branch for help with the preparation of this manuscript, and Ms. Jennevieve Sevilla for excellent technical assistance, and the National Cancer Institute (NCI) for funding our participation in the HVEI panel on medical image perception. We are grateful to the NCI and to Dr. Jeremy Wolfe of Harvard Medical School for organizing the aforementioned RSNA testing facility each year. This study was supported by Army Research Office (ARO) grants W911NF-11-1-0105 and W911NF-15-1-0311 to JH.

References

[1] K. Mueller, et al, "Advancing Value-Based Medicine: Why Integrating Functional Outcomes With Clinical Measures is Critical to Our Health Care Future," J Occup Environ Med, vol. 59, no. 4, pp. e57-e62, 2017.

[2] B. L. Rosenburg, "Quantifying Geographic Variation in Health Care Outcomes in the United States before and after Risk-Adjustment," PLoS One, vol. 11, no. 12, pp. e1066762, 2016.

[3] D. E. Garets & C. M. Garets, The journey never ends: technology's role in helping perfect health care outcomes, Boca Rotan, FL: CRC Press, Taylor & Francis Group, 2016.

[4] D. O. Driscoll, D. Halpenny & M. Guiney, "Radiological error—an early assessment of departmental radiology discrepancy meetings," Ir Med J, vol. 105, pp. 172-174, 2012.

[5] L. J. Grimm, C. M. Kuzmiak, S. V. Ghate, S. C. Yoon, & M. A. Mazurowski, "Radiology resident mammography training: interpretation difficulty and error-making patterns," Acad Radiol, vol. 21, no. 7, pp. 888-892, 2014.

[6] M. A. Mazurowski, H. X. Barnhart, J. A. Baxter, & G. D. Tourassi, , "Identifying Error-making Patterns in Assessment of Mammographic BI-RADS Descriptors among Radiology Residents Using Statistical Pattern Recognition," Acad Radiol, vol. 19, no. 7, pp. 865-871, 2012.

[7] A. Pinto, et al., "Learning from diagnostic errors: A good way to improve education in radiology," Eur J Radiol, vol. 78, no. 3, 372-376.

[8] W. A. Berg, C. Campassi, P. Langenberg, & M. J. Sexton, "Breast Imaging Reporting and Data System Inter- and Intraobserver Variability in Feature Analysis and Final Assessment," AJR Am J Roentgenol, vol. 174, no. 6, 1769-1777, 2000.

[9] L. A. Hardesty, et al., "'Memory effect' in observer performance studies of mammograms," Acad Radiol, vol. 12, no. 3, 286-290.

[10] B. L. Cole, "The handicap of abnormal colour vision," Clin Exp Optom, vol. 87, 258-275, 2004.

[11] M. A. Roubidoux, N. E. Lai, C. Paramagul, L. K. Joynt, & M. A. Helvie, "Mammographic appearance of cancer in the opposite breast: comparison with the first cancer," AJR Am J Roentgenol, vol. 166, no. 1, 29-31, 1996.

[12] M. Y. Sallam & K. W. Bowyer, "Registration and difference analysis of corresponding mammogram images," vol. 3, no. 2, 103-118, 1999.

[13] R. N. Shepard, D. W. Kilpatric, & J. P. Cunningham, "The internal representation of numbers," Cognitive Psychology, vol. 7, no. 1, 82-138, 1975.

[14] R. N. Shepard & S. Chipman, "Second-order isomorphism of internal representations: Shapes of states," Cognitive Psychology, vol. 1, no. 1, 1-17, 1970.

[15] S. Edelman, "Representation is representation of similarities," Behav Brain Sci, vol. 21, no. 4, 449-467, 1998.

[16] N. Kriegeskorte, et al., "Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey," Neuron, vol. 60, no. 6, 1126-1141, 2008.

[17] N. Kriegeskorte, M. Mur, & P. Bandettini, "Representational similarity analysis- connecting the branches of systems neuroscience," Front Syst Neurosci, vol. 24, 2008.

[18] H. O. D. Beeck, J. Wagemans, & R. Vogels, "Inferotemporal neurons represent low-dimensional configurations of parameterized shapes," Nat Neurosci, vol. 4, 1244-1252, 2001.

[19] American College of Radiology, ACR BI-RADS Atlas 5th Edition, Reston, VA: American College of Radiology, 2013.

[20] F. Cutzu & S. Edelman, "Faithful representation of similarities among three-dimensional shapes in human vision," Proc Natl Acad Sci USA, vol. 93, no. 21, 12046-12050, 1996.

[21] F. Cutzu & S. Edelman, "Representation of object similarity in human vision: psychophysics and a computational model," Vision Res, vol. 38, no. 15-16, 2229-2257, 1998.

[22] S. Edelman, "Representation of Similarity in Three-Dimensional Object Discrimination," Neural Comput, vol. 7, no. 2, 408-423, 1995.

[23] S. Edelman & S. Duvdevani-Bar, "Similarity, Connectionism, and the Problem of Representation in Vision," Neural Comput, vol. 9, no. 4, 701-720, 1997.

[24] S. Edelman & S. Duvdevani-Bar, "A model of visual recognition and categorization," Philos Trans R Soc Lond B Biol Sci, vol. 352, no. 1358, 1191-1202, 1997.

[25] W. A. Berg & W. T. Yang, Diagnostic imaging. Breast. 2nd Edition, Salt Lake City: UT, 2014.

[26] J. Sevilla & J. Hegdé, "Deep Visual Patterns Are Informative to Practicing Radiologists in Mammograms in Diagnostic Tasks," Journal of Vision, vol. 17, no. 90, 614, 2017.

[27] E. Bart & J. Hegdé, "Deep Synthesis of Realistic Medical Images: A Proof-of-Principle Study," Frontiers in Neuroinformatics, submitted 2018.

Authors Biographies

Jay Hegdé is a computational neuroscientist. He obtained his PhD in molecular biology from the University of Rochester. He switched to computational neuroscience during his post-doctoral studies, and obtained training in neurobiology, awake-behaving monkey neurophysiology, cognitive psychology, and machine learning. He is primarily interested in understanding how the brain works, and studies brain function in human subjects and monkeys. His current research focuses on understanding brain function and dysfunction under real-world conditions.

Evgeniy (Eugene) Bart is a computer scientist. He obtained his PhD from the Weizmann Institute in Israel, and obtained post-doctoral training in computer science and machine learning at the University of Minnesota, and Caltech. His research interests are in machine vision, machine learning and computer science.

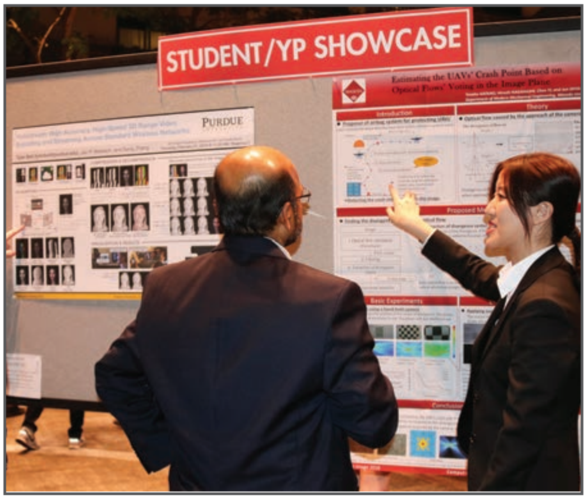
JOIN US AT THE NEXT EI!

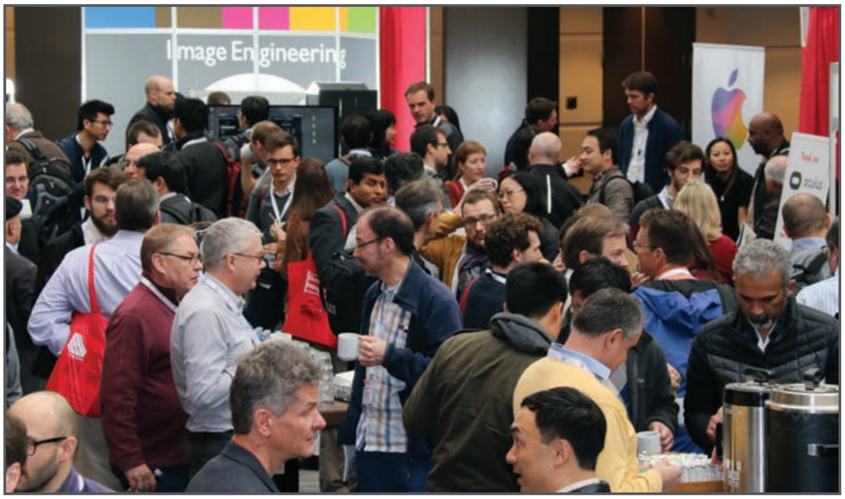
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





- AR is a fundamental and transformative interface and platform.
- 20+ years from my initial working demo to today's AR displays
- Optimistic that it won't take another 20 to reach ubiquitous consumer AR!

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org