https://doi.org/10.2352/ISSN.2470-1173.2021.10.IPAS-247

Breast Cancer Tissue Sub-Region Classification from Second Harmonic Generation Imagery via Machine Learning

Wencheng Wu, Beilei Xu, Edgar A. Bernal; Rochester Data Science Consortium, University of Rochester; Rochester, NY, USA
Robert L. Hill; Harmonigenic Inc.; Rochester, NY, USA
Danielle Desa, Edward B. Brown; Dept. of Biomedical Engineering, University of Rochester; Rochester, NY, USA

Abstract

Cancer treatment involves complex decision-making processes. A better understanding of recurrence risk at diagnosis, as well as prediction of treatment response (e.g., to choose the most cost-effective treatment path in an informed manner) are needed to produce the best possible outcomes at the patient level. Prior work shows that the forward/backward (F/B) ratio calculated from Second Harmonic Generation (SHG) imagery can be indicative of risk of recurrence if relevant tissue sub-regions are selected. The choice of which sub-regions to image is currently made by human experts, which is subjective and labor intensive. In this paper, we investigate machine learning methods to automatically identify tissue sub-regions that are most relevant to the prediction of breast cancer recurrence. We formulate the task as a multi-class classification problem and use support vector machine (SVM) classifiers as the inference engine. Given the limited amount of data available, we focus on exploring the feature extraction stage. To that end, we evaluate methods leveraging handcrafted features, deep features extracted from pre-trained models, as well as features extracted via transfer learning. The results show a steady trend of improvement on the classification accuracy as the features become more data-driven and customized to the task at hand. This is an indication that having larger amounts of labeled data could be beneficial for improving automated methods of classification. The best results achieved, using features learned via transfer learning from ResNet-101, correspond to 85% accuracy in a 3-class problem and 94% accuracy in a binary classification problem.

I. Introduction

Cancer treatment involves a series of complex decision-making processes. For solid tumors, treatment guidelines have evolved from surgical intervention alone, to surgical treatment followed by adjuvant chemotherapy or just endocrine therapy for early stage cancers with low risk of recurrence, to neoadjuvant chemotherapy where chemotherapy is given before surgery for cancers with a high risk of distant metastatic recurrence or to shrink an otherwise inoperable tumor. A better understanding of recurrence risk at diagnosis as well as prediction of treatment response to choose the most cost-effective treatment strategy are needed in order to produce the best possible treatment outcomes. Consequently, methods predicting recurrence risk at diagnosis with improved risk stratification for cancer patients are required in order to better identify early treatment options for patients. In [1], an optical diagnostic assay technology that leverages an intrinsic optical signature from collagen in standard diagnostic tissue slides is discussed. The predictor is derived from the average pixel-intensity ratios of the forward (F) to backward (B) second harmonic generation (SHG) light scatter images, denoted F/B ratio. It was further shown in [2] that the prediction accuracy using the SHG F/B ratio depends on the tissue sub-region(s) involved in the computation. It is thus critical to determine the most relevant sub-regions that are suitable to compute the F/B measurements, where relevance is measured in terms of the accuracy of the derived risk prediction models.

In this paper, we propose methods for automatically predicting cancer recurrence from SHG imagery, as discussed conceptually in [1] and [2]. Fig. 1 illustrates a high-level view of the process. Current practices involve a human expert selecting relevant regions in the SHG images of tissue samples. The resulting forward and backward image-pair is analyzed to yield a biomarker, namely the averaged F/B ratio, computed as per the image analysis process discussed in [1]. The F/B ratio is shown to be indicative of the risk of recurrence, and is used to categorize patients into high- or low-risk categories. The associated risk over time can be estimated based on the Kaplan Meier Curve (KMC) [3] derived from clinical studies such as those in [1]. There is ample room for improvement across the multiple steps involved in the process. For example, the image analysis stage requires pixel segmentation highlighting collagen for accurate F/B ratio computation. This step is currently done by manually selecting thresholds for the F and B image-pairs. Advances in image processing and computer vision in the field of image segmentation [4]-[7] can be leveraged to automate this step. Another step that can be improved with automation is the biomarker discovery and selection, which currently limits the granularity of the recurrence risk prediction (e.g., high vs. low or use of group KMC for individual risk estimation). This constraint is the result of the limited availability of patient data in clinical studies, as well as the low-dimensionality of the F/B ratio feature. To improve on this aspect, research in data-driven biomarker discovery based on learned image features can potentially improve the richness of the descriptors, which in turn would allow better prediction of individual recurrence risks. As higher dimensional features are incorporated, recent advances in deep survival analysis [8][9] may be well suited for predicting individual risk.

In this paper, we focus on exploring steps towards automating the process of identifying relevant sub-regions, as discussed in [2], via machine learning. This corresponds to the automation of the first step of the process depicted in Fig. 1. The main motivation lies in our preliminary findings on the importance of selecting image regions with greater diagnostic power. Additional benefits that accompany the deployment of automated decision-making processes, namely improved accuracy, repeatability and

processing throughput, and decreased subjectivity, are also intended byproducts. We expect the automation to be particularly beneficial in this application, as it will enable the efficient acquisition of more relevant samples, which should in turn improve the prediction power of the process and close the loop on the SHG imaging and diagnosis process.

The rest of this paper is organized as follows: Section II introduces the problem that our proposed methodology intends to solve and a brief discussion of prior work. Section III provides quantitative and qualitative evaluation results and experimental justification of the efficacy of the algorithm. Conclusion and future work are presented in Section IV.

II. Problem Statement and Related Work

In this section, we briefly describe the problem and discuss some of the relevant prior work in the field. As discussed earlier, being able to identify relevant regions of interest (ROIs), i.e., tissue sub-regions, in a sample is a critical step in medical diagnosis. This step is often performed manually by trained pathologists in a process that can be time-consuming and subjective. In [10], researchers formulated the ROI identification process as an unsupervised learning problem and solved it via phenotyping. To that end, randomly selected ROIs were clustered into groups corresponding to different phenotypes, and the effectiveness of the resulting phenotypes was measured based on how well they predicted patient outcomes. The benefits of this approach are: (1) no labeling of relevant ROIs is needed since the selection is done randomly; and (2) new discoveries of relevant phenotypes is possible. The disadvantages are: (1) a large number of samples with patient outcomes is needed; (2) the process does not directly measure the effectiveness of the feature extraction and clustering stages; and (3) the methodology may not be easily interpretable as it is somewhat data-driven instead of heuristics-based. Due to the limitation in the number of available samples and motivated by the need for explainability and interpretability, we focus on investigating machine learning methods for localizing ROIs related to tissue sub-regions already identified by human experts as the most relevant to the prediction of breast cancer recurrence. By doing so, our methods will yield regions that match the experts' knowledge of relevance and are thus inherently explainable. Also, since we aim at addressing an altogether simpler task, the data requirements are more modest. From the perspective of image classification research, the best community practices involve the use of end-to-end image classification frameworks via deep learning architectures. This approach, however, is extremely data hungry. In the context of this study, and due to the scarcity of labeled samples, we address the feature extraction and inference stages independently. More specifically, we formulate the task as a multi-class classification problem and use the support vector machine (SVM) classifiers [11] as the inference engine, while performing thorough exploration of the feature extraction stage. We choose the linear kernel SVM as the common classifier in our study due to its simplicity and proven track record. We systematically investigate two aspects of feature extraction: the degree of feature learning (from hand-engineered to data-driven) and the extent of fusion between the F and B image components. To understand the impact of the degree of feature learning on classification accuracy, we study three types of features with increasing levels of learning: hand-crafted (e.g., HOG [12])), pre-trained deep features (e.g., AlexNet [13] and ResNet-101 [14] trained on ImageNet [15]), and refined deep features through transfer learning of a pre-trained network. This approach allows us to test the efficacy of features as a function of their relevance to the dataset at hand, and to elicit conjectures about the hypothetical behavior of a trained network with larger amounts of labeled data. Since there are F and B image components available for each ROI, we also explore the classification performance using different combinations thereof. That is, to understand the impact of fusion among F and B image pairs on the classification accuracy, we study the effect of using the F and B components separately, as well as a fused version of both components towards feature extraction. Fusion schemes implemented include early fusion by (1) feature concatenation of HOG and pre-trained deep features extracted from both image components, and (2) use of pseudo RGB images resulting from stacking of the available components (e.g., [F|B|F]) as inputs to a CNN. A brief description of the features explored is included in Table 1. Through these studies, our objectives are to:

- understand whether there is observable performance improvement as the degree of feature learning increases, as well as validate whether machine learning is feasible with current limitation of data;
- understand the feasibility of automating sub-region selection with computer vision techniques and enable an efficient SHG imaging platform where either the samples are densely imaged and then selectively processed (e.g., as in [10]), or where the image regions are first pre-scanned at low resolution and then selectively imaged with higher quality (i.e., a closed-loop imaging platform);
- develop a more effective strategy for sub-region selection involving combinations of the F and B SHG image components.

Table 1. Brief description of features considered.

Group	Notation	Brief Descriprion
A	HOG(F)	HOG extracted from F images
	HOG(B)	HOG extracted from B images
	HOG(FB)	Concatenation of HOG(F) & HOG(B)
B	AlexNet(F)	Pre-trained AlexNet feature extracted from F images
	AlexNet(B)	Pre-trained AlexNet feature extracted from B images
	AlexNet(FB)	Concatenation of AlexNet(F) & AlexNet(B)
	AlexNet(F|B|F)	Pre-trained AlexNet feature extracted from (F,B,F) pseudo-RGB images
C	ResNet(F)	Pre-trained ResNet feature extracted from F images
	ResNet(B)	Pre-trained ResNet feature extracted from B images
	ResNet(FB)	Concatenation of ResNet(F) & ResNet(B)
	ResNet(F|B|F)	Pre-trained ResNet feature extracted from (F,B,F) pseudo-RGB images
D	Refined ResNet(F)	Refined ResNet feature via transfer learning on F images

III. Experiments and Results

The following sections describe the experimental framework employed as we evaluate the feasibility of the proposed approach. Due to the scarcity of patient samples, we conducted two types of experiments, one quantitative and one aimed at testing the robustness of the proposed approach.

Quantitative Study

The dataset for this experiment, DataSet#1, consists of tissue samples from 96 estrogen receptor-positive (ER+) breast cancer patients. Three tissue sub-regions, namely interface (stroma next to cancer cell bulk regions), bulk (mostly cancer cells with some

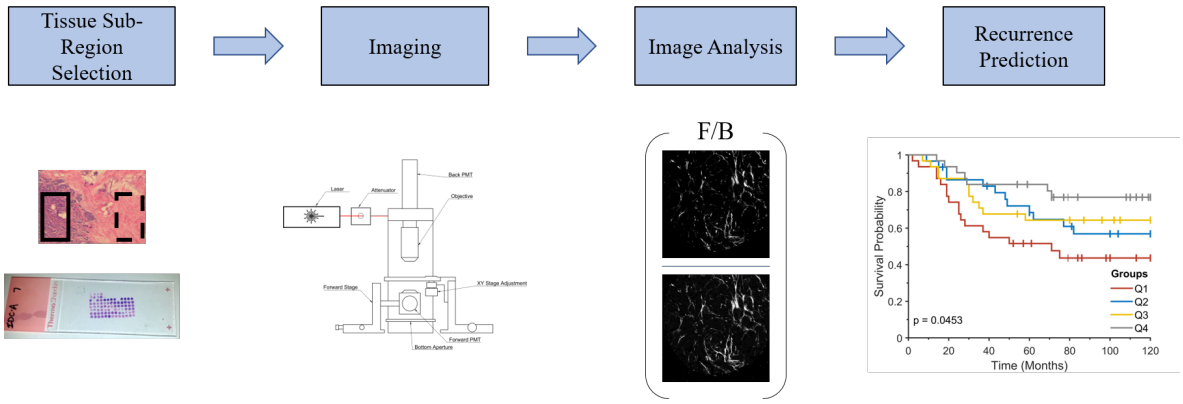


Figure 1. *System flowchart for cancer recurrence prediction using SHG image analysis.*

stroma), and far (far away from cancer cells including normal breast tissue), are identified and imaged using SHG (see Fig. 2). Whenever possible, all three sub-regions are selected and imaged from each sample. In total, 795 F and B image pairs are available: 297 interface, 297 bulk, and 171 far. In [2], it was shown that the relevance of the biomarkers to the prediction of breast cancer recurrence and treatment response depends on the region of tissue they are extracted from, with the most discriminative being those features extracted from the interface, followed by tissue in the bulk and lastly the far regions. With this in mind, our goal is to differentiate the three sub-regions with high accuracy (i.e., in the form of a 3-class classification formulation), or to discriminate the far regions from the interface and bulk regions (binary or 2-class classification problem).

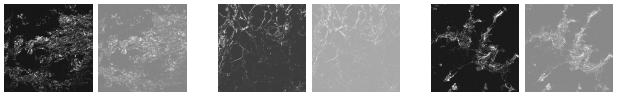


Figure 2. *DataSet#1: Example F (left) and B (right) image pairs from SHG imaging of the tissue of a patient from (a) interface, (b) bulk, and (c) far regions. Images have been enhanced for illustration purposes.*

Fig. 3 summarizes the results of the multi-class classification task. We evaluated each of the methods listed in Table 1 a total of 10 times using random training/test splits of 80%/20%; the figure includes boxplots of the performance on the test set. We performed experiments using different combinations of F and B imagery. The label F (B) refers to using features extracted from the F (B) image only. The label FB refers to the concatenation of F and B derived features. The label F|B|F refers to features extracted from pseudo-RGB images constructed by stacking F, B, F images into the Red, Green and Blue channels, respectively, of an RGB image. We chose to duplicate the F plane since it showcases better contrast than the B component. It can be seen that the best results are achieved by the method with higher degree of learned features, namely feature group D, with 85% accuracy in the 3-class scenario and 94% accuracy in the binary classification task. Two interesting observations can be drawn from the figure: (1) using F imagery alone yields more accurate and robust classification; and (2) the more relevant learning is involved in the feature extraction (from left to right), the better results can be achieved. The first observation aligns with the fact that F images of our tissue samples exhibit better contrast and thus have better signal to noise ratio (SNR). When training data is limited, combining F image with low SNR B image does not improve the performance. However, we expect the result from F&B images would improve over individual F or B images if large training data set is used. The second observation is an indication that having more labeled data in the future (so that a higher degree of feature learning can be carried out) would be beneficial to the performance of the automated method. This would be part of future research.

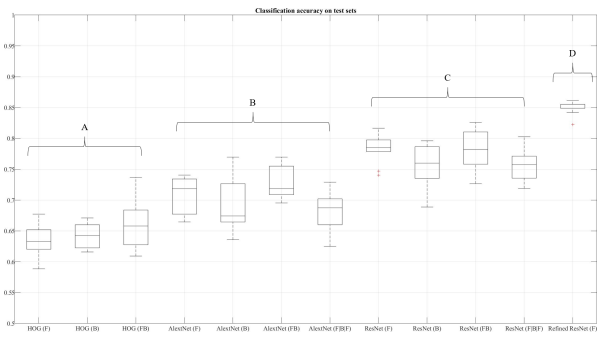


Figure 3. *Results of multi-class classification using an SVM classifier operating on different features.*

Robustness Study

In this experiment, we utilize a portion of the dataset in [1] to validate our method from the perspective of generalizability and robustness. This dataset, DataSet#2, consists of 344 human breast tumor samples from 125 ER+ breast cancer patients. This data comes from a collection at the Erasmus Medical Center (Rotterdam, Netherlands), primarily extracted from one breast cancer genetic expression study [16]. Fig. 4 illustrates examples of SHG F&B image pairs from three patients. Note that, since the regions imaged in these samples were intentionally selected to cover the tumor, we expect the center of the F&B images to be located in the bulk region defined in the quantitative experiment. Depending on the size of the tumor, the outer regions may cover some interface-region as discussed earlier, but this is not guaranteed. Due to this limitation, we only use the data to validate how well our methods can classify bulk-region samples. Another limiting characteristic

of these samples is that the samples are prepared and fixed on a disk; as a result, only circular potion at the center of the images is considered to be valid (i.e., corresponding to tissue rather than background). Compared to DataSet#1 (see Figs. 2 and 4) where images are relevant across the entire field of view, the distinct border and the outer background pixels can potentially cause issues with traditional classification models. At the same time, the novel data characteristics (that is, those not seen by the model during training) allow us to validate the robustness of the models trained on DataSet#1. To sum up, the F&B images in DataSet#2 are only weakly labeled with mostly bulk-region tissue (partially contaminated by interface and background pixels) and have novel features not necessarily seen by the model during training. With these issues in mind, we only use this dataset to test the robustness of the methods.

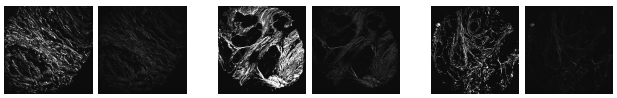


Figure 4. *DataSet#2: Example F (left) and B (right) image pairs from SHG imaging of human breast tumor samples from three patients. Images have been enhanced for illustration purposes.*

To further validate the models trained on DataSet#1, we apply the first and the last trained models listed in Table 1, namely those based on HOG(F) and Refined ResNet(F), to classify samples in DataSet#2. The results are shown in Fig. 5 which summarizes the results of applying the two aforementioned models to the center 512×512-pixel portion and to the entire 1024×1024-pixel images in DataSet#2. As discussed, we expect the center portion to contain mostly bulk-region tissue (depending on the size of the tumor) and to be largely devoid of the unwanted background pixels. Comparing Figs. 5 (a)(b) and (c)(d), it can be seen that the classifier based on Refined ResNet(F) is much more accurate at the task than the HOG(F) classifier. Since the bulk region can vary from sample to sample in this dataset due to tumor size variation, a better indication of the model's accuracy might be the total accuracy of bulk+interface. Using this criterion, our Refined ResNet(F) can achieve 90% accuracy for DataSet#2 in the binary case (i.e., bulk or interface vs. far), which is on par with the test accuracy for DataSet#1. This shows that Refined ResNet(F) has better generalizability beyond the samples it saw at training. Comparing Figs. 5a and 5b, it is apparent that Refined ResNet(F) is very robust against the unobserved image characteristics (distinct border and outer background pixels not present in the training set). Lastly, comparing Figs. 5c and 5d, it can be concluded that the performance of HOG(F), on the other hand, seems to be affected by these disturbances. The results are not surprising since HOG features are known to be susceptible to variations due to translation, scaling, rotation, and changes in the background as they focus mostly on local features. Deep features such as ResNet, on the other hand, incorporate multi-scale (local and global), hierarchical features and are more robust against a wider range of factors of variability.

In summary, our experiments show that: (1) it is sufficient to use only Forward images for tissue sub-region classification especially when the training data is limited; (2) with a small amount of samples for transfer learning, Refined ResNet(F) is able to

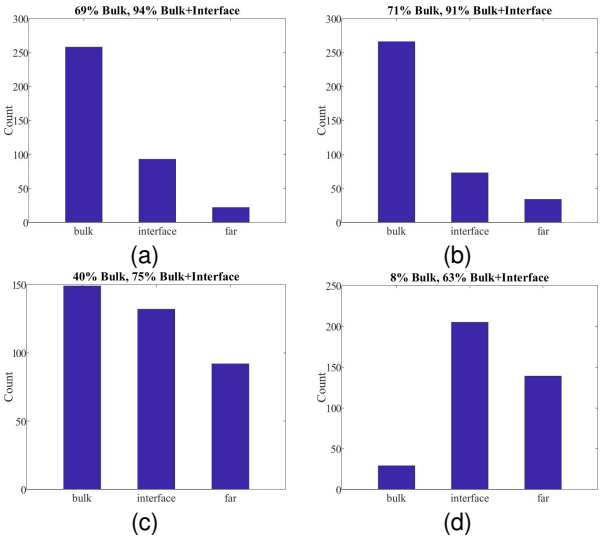


Figure 5. *Classification results on DataSet#2 using Refined ResNet(F) on (a) the center portion and (b) the entire image, and using HOG(F) on (c) the center portion and (d) the entire image.*

achieve high accuracy; and, (3) Refined ResNet(F) is more robust and generalizes better than Hog(F). These results seem to indicate that increasing the degree of feature learning would greatly benefit the task of automatically localizing relevant tissue sub-regions for breast cancer recurrence risk prediction.

IV. Conclusions and Future work

In this study, we investigated machine learning methods for classifying sub-regions of tissue from SHG imagery. We focused on the feature extraction stage and found quantitatively and qualitatively that the more closely related the feature extraction is to the task at hand, the better is the classification accuracy. This indicates that as a larger amount of labeled data become available, increased accuracy can be expected from the resulting tissue sub-region classification algorithm, particularly if the algorithm in question is deep in nature. Our ultimate goal is to develop a system where tissue samples are first densely imaged for analysis, an algorithm identifies regions of interest, and the most relevant regions are selected for further analysis; we expect this informed triaging will enable more accurate recurrence prediction. The work presented herein is the first step towards that goal.

Acknowledgments

This work is partially supported by a grant from the Center of Excellence in Data Science (CoE), Goergen Institute for Data Science, University of Rochester, funded by the New York State Department of Economic Development.

References

[1] K. Burke, M. Smid, R. P. Dawes, M. A. Timmermans, P. Salzman, C. H. M. van Deurzen, David G. Beer, J. A. Foekens and E. Brown, "Using second harmonic generation to predict patient outcome in solid tumors," BMC Cancer, 15:929, 2015.

[2] Desa D, Turner B, Buscaglia B, Hill R, Majeski J, Choe R, Strawderman R, Kuo C, Hicks D, Brown E. "Using multiphoton laser-scanning microscopy to assess neoadjuvant therapy outcome in core

needle biopsies: a novel methodology," San Antonio Breast Cancer Conference. December 2018.

[3] E. L. Kaplan, Paul Meier, "Nonparametric Estimation from Incomplete Observations," Journal of the American Statistical Association, 53:282, 457-481, 1958. doi: 10.1080/01621459.1958.10501452.

[4] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms." IEEE Transactions on Systems, Man, and Cybernetics. 9(1) , pp. 62–66, 1797.

[5] D. Bradley, G. Roth, "Adapting Thresholding Using the Integral Image," Journal of Graphics Tools. 12(2), pp.13–21, 2007.

[6] M. Seyedhosseini, M. Sajjadi and T. Tasdizen, "Image Segmentation with Cascaded Hierarchical Models and Logistic Disjunctive Normal Networks," 2013 IEEE International Conference on Computer Vision, Sydney, NSW, pp. 2168-2175, 2013, doi: 10.1109/ICCV.2013.269.

[7] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," In: Stoyanov D. et al. (eds) Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2018, ML-CDS 2018. Lecture Notes in Computer Science, vol 11045. Springer, Cham.

[8] Lee, C. et al. "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks." AAAI (2018).

[9] Milad Zafar Nezhad, Najibesadat Sadati, Kai Yang, Dongxiao Zhu, "A Deep Active Survival Analysis approach for precision treatment recommendations: Application of prostate cancer," Expert Systems with Applications, Volume 115, pp. 16-26, 2019.

[10] X. Zhu, J. Yao, F. Zhu and J. Huang, "WSISA: Making Survival Prediction from Whole Slide Histopathological Images," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 6855–6863, 2017.

[11] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," in Machine Learning 20(3), pp. 273-297, 1995.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, pp. 886-893 vol. 1, 2005, doi: 10.1109/CVPR.2005.177.

[13] Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems 25, pp. 1097-1105, 2012.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[15] ImageNet. http://www.image-net.org

[16] Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. "Geneexpression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," Lancet. 365(9460), pp. 671-679, 2005.

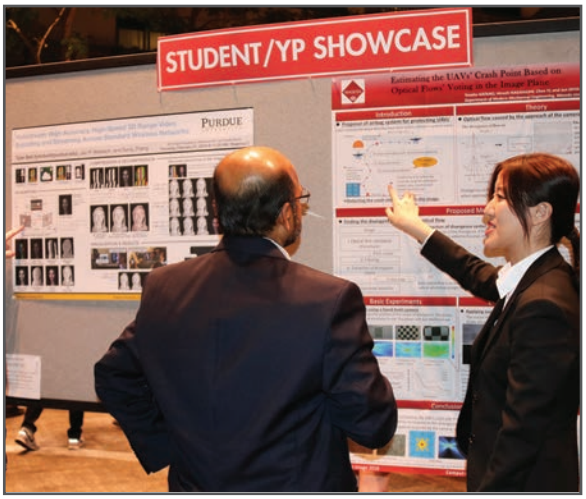
JOIN US AT THE NEXT EI!

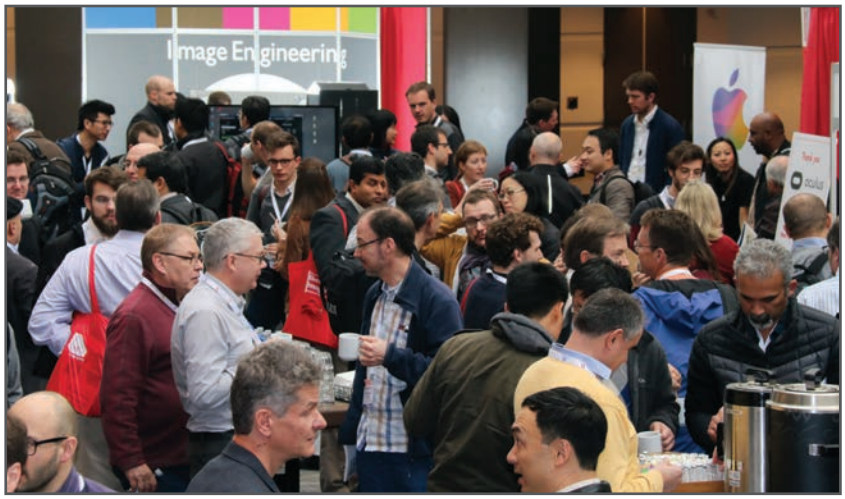
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

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





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

www.electronicimaging.org

IS&T
imaging.org