https://doi.org/10.2352/ISSN.2470-1173.2019.11.IPAS-267

Construction of Facial Emotion Database Through Subjective Experiments and Its Application to Deep Learning-Based Facial Image Processing

Tomoyuki Takanashi, Keita Hirai, Takahiko Horiuchi
Department of Imaging Sciences, Chiba University, 1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522 Japan.

Abstract

As the development of interactive robots and machines, studies to understand and reproduce facial emotions by computers have become important research areas. For achieving this goal, several deep learning-based facial image analysis and synthesis techniques recently have been proposed. However, there are difficulties in the construction of facial image dataset having accurate emotion tags (annotations, metadata), because such emotion tags significantly depend on human perception and cognition. In this study, we constructed facial image dataset having accurate emotion tags through subjective experiments. First, based on image retrieval using the emotion terms, we collected more than 1,600,000 facial images from SNS. Next, based on a face detection image processing, we obtained approximately 380,000 facial region images as "big data." Then, through subjective experiments, we manually checked the facial expression and the corresponding emotion tags of the facial regions. Finally, we achieved approximately 5,500 facial images having accurate emotion tags as "good data." For validating our facial image dataset in deep learning-based facial image analysis and synthesis, we applied our dataset to CNN-based facial emotion recognition and GAN-based facial emotion reconstruction. Through these experiments, we confirmed the feasibility of our facial image dataset in deep learning-based emotion recognition and reconstruction.

Introduction

Understanding emotions in facial expressions is important in human communication. In addition, as the development of interactive robots and machines such as Pepper and NAO [1], studies to recognize and reconstruct facial emotions have become important research areas. For example, Kalegina et al. [2] surveyed that specific features affects perception of the rendered robot faces.

One of the approaches to recognize and understand facial emotions is to use facial image features. For example, Nomiya et al. [3] proposed a method to construct facial emotion recognition models by extracting feature points related to eyebrows, eyes, nose, and mouth from facial images. Bukar et al. [4] proposed an age change algorithm based on facial features for searching missing people around the world. Another approach on facial emotion recognitions is based deep learning which has been recently developed. For example, Alizadeh et al. [5] developed a facial emotion estimator based on convolutional neural network (CNN) which is one of the deep neural network. They used Kaggle which provided approximately 37,000 grayscale facial images tagged with 7 kinds of emotions for implementing a facial emotion estimator. Liao et al. [6] proposed a method to transfer visual attributes (color, tone, texture, style, etc.) of an input facial image to another target facial image using CNN.

In the researches based on deep learning as described above, big data of facial emotion images or expression image dataset are generally required. Big data is a dataset including large amount of information. In recent years, researches on collecting big data by crowdsourcing are also actively studied [7]. For constructing big data, it is a common approach to collect data from web site such as social media, social networking service (SNS), and image retrieval systems. When collecting big data of facial emotion images, such images are retrieved based on keyword search systems using emotion terms. However, the big data collected by such a method includes a large amount of noise data. For example, facial image big data based on the above keyword search includes noises such as facial images with wrong emotion tagged or obviously non-facial images. Also, several research groups have already published facial image datasets with emotion tag. However, these datasets have some barriers such as download fees, wrong emotion tags or number of facial images.

For constructing a dataset with reliability and universality, it is necessary to collect facial images from a lot of people with various facial expressions. For achieving this goal, we built facial emotion image dataset instead of existed image data set. When employing the above data collecting approach by web site and keyword search, the dataset includes a large amount of noise data. In this case, deep learning-based image processing using such image dataset will not provide good result. Therefore, we focused on "good data" which is a dataset with less noise. Also, we consider that even if the number of good data is smaller than that of big data, the good data provides better result compared with big data. Therefore, in this research, we confirm the feasibility of the facial emotion image dataset based on good data by comparing with big data. First, we collect facial expression image big data from SNS. Next, we conduct subjective experiments for selecting good data from big data. Then, we validate our dataset in deep learning-based facial image processing such as facial emotion recognition and reconstruction. In the facial emotion recognition, we construct two kind of emotion recognition networks using big data and good data. Then we compare the accuracy between the emotion recognition networks. In the facial emotion reconstruction, we applied our good data to StarGAN [8] for manipulating facial emotion.

Construction of facial emotion image dataset

In the research field of cognitive psychology, various studies on facial emotion recognition have been investigated. Ekman et al. [9] suggested that expressions for specific emotions are universal regardless of differences in languages and cultures. In addition, Ekman et al. said that the basic emotions are classified into six

types: ANGER, DISGUST, FEAR, HAPPINESS, SADNESS and SURPRISE.

In this research, we construct two types of datasets that store facial emotion images with one tag of seven categories (six emotions proposed by Ekman et al.; ANGER, DISGUST, FEAR, HAPPINESS, SADNESS, SURPRISE, and NEUTRAL). Two types of facial image dataset are big data and good data. Figure 1 shows the procedure for collecting big data and good data. Figure 2 shows examples of emotions.

Big data collection using SNS

In this research, Twitter, one of the representative SNSs, was used for collecting facial images. The reason is that Twitter has the following features and is suitable for the big data collection and analysis.

- Twitter has 328 million monthly active users. (as of April 2017) [10]. A large number of sentences and image data are constantly posted every day.
- Everyone can access Twitter's tweets. We can also collect only tweets and data under specific conditions by using Twitter Application Programming Interface (API).

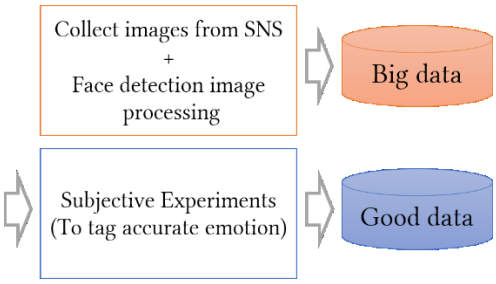


Figure 1. Procedure for collecting big data and good data.

(a) ANGER *(b) DISGUST* *(c) FEAR*

(d) HAPPINESS *(e) SADNESS* *(f) SURPRISE*

(g) NEUTRAL

Figure 2. Examples of emotions. (The original of the face image being used is PAKUTASO(www.pakutaso.com))

A flowchart of the facial expression image dataset construction based on Twitter is shown in Fig. 3. The details are as follows:

1. Collecting images by searching tweets with emotional terms. We implemented an automatic tweet collection system using Twitter API [11]. With this system, we search for tweets with images containing the emotion terms shown in Table 1 and collect up to 100 images per hour for each term.
2. Extracting facial area of images and tag emotion terms to the facial images. We used the OpenCV [12] Cascade-Classifier to detect face regions. The extracted face area is temporarily stored as a face image and tagged using emotional terms which are used in the keyword search with Twitter API.
3. Deleting non-facial images that were misextracted by the CascafeClassifier. We delete non-facial images according to the following criteria.
   - Obviously non-skin tone images: Checking the HSV value of each pixel. We assumed skin tone are within $(0^\circ \leq H \leq 60^\circ)$ & $(0 \leq S \leq 40)$. If 50% of the whole pixels in an image included the skin tone, we regarded the image as a facial image. Other images were deleted.
   - Store the remaining images in the facial expression image dataset as "big data."
4. Saving the remaining image in the facial emotion image dataset.

However, noise image data as shown in Figure 4 are remained when using the method in the above steps 2 and 3. In addition, although a face image is tagged using a search query of an emotion term, a lot of facial images are incorrectly tagged (a lot of facial images with wrong emotion terms).

The data collected in the above procedure is big data with large amounts of data and noise. In the next section, we refine the data based on the subjective experiment and collect good data (facial images with accurate emotion terms).

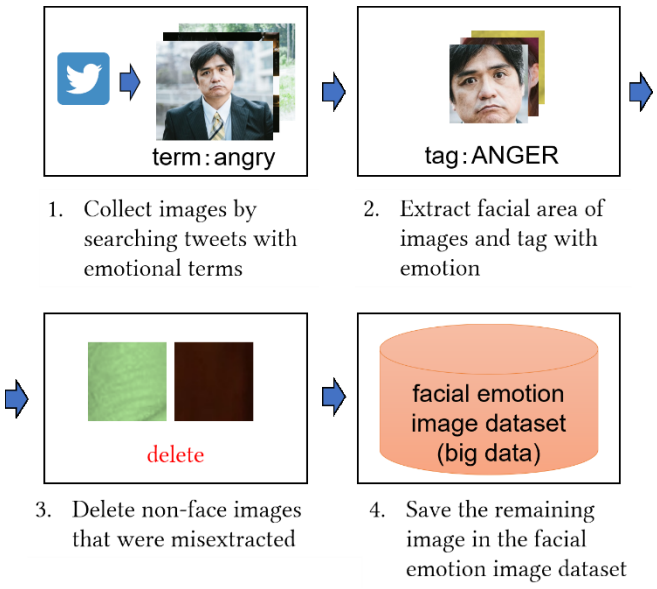


Figure 3. The flowchart of the facial expression image dataset construction based on Twitter. (The original of the face image being used is PAKUTASO(www.pakutaso.com))

Table 1. Emotion terms by emotion categories.

Emotion category	Emotion terms
ANGER	anger, angry, annoy, furious, fury, mad, rage
DISGUST	disgust, disgusted, hate, hated, aversion, dislike
FEAR	fear, fearful, afraid, awe, horror, panic, scare, terror
HAPPINESS	happiness, happy
SADNESS	sadness, sad, distress, grief, lament, sorrow
SURPRISE	surprise, surprised, amazement, fright, shock, surprising, wonder
NEUTRAL	(We didn't collect by term search)

Figure 4. Noise data close to the face image.

Good data selection based on subjective evaluation experiment

As described above, the big data has a huge amount of noise data. In other words, the big data includes a lot of facial images with wrong emotion tags. Therefore, we deleted noise data through subjective experiments and selected only good data. We also divided the good data into two types: training data and test data for deep learning-based image processing.

Selection of good data.

We conducted subjective evaluation experiments with the following procedure for collecting the good data.

1. Present a facial image of big data and its emotion tag to one subject. (One facial image was presented to one subject.)
2. If the subject evaluates the emotion tag as incorrect, the facial image is not used as the good data.
3. Present an image with an incorrect emotion tag. In this situation, subject answers whether the emotion is NEUTRAL or not for collecting good data of NEUTRAL, because it is difficult to collect facial images with neutral emotion based on the keyword search using Twitter API.

One facial image was evaluated by one subject. For collecting good data, 8 subjects participated to the subjective evaluation experiment. This work took about 1 month.

Table 2 shows the number of each data. In the table, the column of (1) shows the number of images that was collected by the keyword search with emotion terms. The column of (2) shows the number of the big data that was collected through the face detection image processing and the noise deleting process. The column of (3) shows the number of the good data that was collected by the subjective experiment.

Division of good data to training and test data.

In this study, the good data was collected for deep-learning-based image processing such as CNN-based emotion recognition and GAN-based emotion reconstruction. Good data is then used as training data and test data in such deep learnings. We acquire more accurate test data from good data through subjective evaluation experiment as follows.

1. Randomly select facial emotion images from the good data.
2. Present facial emotion images to multiple subjects. They answered one emotion term from seven kinds of emotions (ANGER, DISGUST, FEAR, HAPPINES, SADNESS, SURPRISE, or NEUTRAL).
3. Facial emotion images that are recognized as one specific emotion by the multiple subjects is stored as the test data.

We conducted three subjective experiments by eight subjects. The first and second times were conducted to collect test data for all seven emotion categories. The third time was done to collect test data of FEAR category which was insufficient through only the first and second experiments. Table 3 shows the result of subjective experiments for collecting test data. In (X/Y) of the table, Y is the total number of subjects, and X is the threshold of how many subjects select the same one from the seven emotions. For example, (7/8) means that seven of eight subjects watched a certain image and selected the same one emotion. Then the certain image becomes test data. Essentially, the threshold value was set to seven subjects. However, because it was difficult to collect the certain images of DISGUST and FEAR under the condition of (7/8), the threshold of DISGUST was set to six, and the one of FEAR was set to four or three subjects. The dark color cells in the table show non-used test data.

Table 2. The number of each data.

Emotion category	(1) Images from Twitter	(2) Big data	(3) Good data
ANGER	310,080	45,450	1,112
DISGUST	141,525	35,228	784
FEAR	427,533	103,362	738
HAPPINESS	264,093	10,519	1,042
SADNESS	180,926	41,987	690
SURPRISE	358,693	50,809	549
NEUTRAL	-	-	561
Total	1,682,850	383,299	5,476

Table 3. Results of subjective experiments for test data.

The number of good data chosen randomly	First time	Second time	Third time	TOTAL
	701	856	200	1757
ANGER	12(7/8)	30(7/8)	20(3/8)	42
DISGUST	24(6/8)	17(6/8)	24(3/8)	41
FEAR	19(4/8)	11(4/8)	11(3/8)	41
HAPPINESS	93(7/8)	13(7/8)	6(3/8)	106
SADNESS	34(7/8)	99(7/8)	37(3/8)	133
SURPRISE	47(7/8)	26(7/8)	16(3/8)	73
NEUTRAL	75(7/8)	23(7/8)	88(3/8)	98

Table 4. The number of training data and test data.

Emotion category	Good data	Training data	Test data
ANGER	1,112	1,050	42
DISGUST	784	730	41
FEAR	738	653	41
HAPPINESS	1,042	949	106
SADNESS	690	554	133
SURPRISE	549	502	73
NEUTRAL	561	504	98
Total	5,476	4,942	534

For the training data, we used the facial emotion images that were not selected from the good data in the subjective experiments or were not adopted in the subjective experiments. Table 4 shows the number of the training data and the test data.

Verification of facial emotion recognition using CNN

In order to confirm the usefulness of the good data in deep learning compared to the big data, we trained both of them in CNN which was a type of deep neural network and created CNN models that performed emotion classification of six emotions excluding NEUTRAL. (Because we did not collect big data of NEUTRAL. Results of seven-category emotion classification using good data as training data will be described later.)

Data set

Table 5 shows the number of the big data and the good data used as the training data and the number of test data. These numbers were same for each category. The big data used as training data in big data was extracted from Table 2.

Network architecture

We used Caffe [13] for deep learning framework. CNN's network architecture is the same as that of CaffeNet [14]. The trainings were executed in scratch. The input images were a square facial expression images which were resized to 227 × 227. Table 6 shows the hyper-parameter of CNN training.

Results

Figure 5 shows results of 6 category classification by the networks trained with big data and good data. Estimated accuracy of the networks trained with the big data and the good data were 30% and about 57%, respectively. From these results, we confirmed that the good data provided a classification model with high accuracy, even though the number of the training images of good data was one-twentieth the number of big data. Figure 6 shows the result of 7 category classifications by the network trained with the good data. Estimated accuracy was about 51%. From Figure 6, the accuracy of HAPPINESS and SURPRISE is high. However, in other categories, the accuracies are low. The following are conceivable reasons:

- The number of training data is small.
- A lot of facial emotion images tagged as FEAR consists of faces with opened mouth. However, it is judged as SURPRISE because of the opened mouth.
- A lot of facial emotion images tagged as DISGUST and NEUTRAL are similar in point of eyebrows shape.
- A lot of facial emotion images tagged as DISGUST and FEAR are similar in point of mouse shape.

Table 5. The number of big data and good data used as training data and the number of test data when training six-category classification. (Numbers enclosed in parentheses are numbers of data when training seven-category classification.)

Emotion category	Training data		Test data
	Big data	Good data	
ANGER	10,000	500	20
DISGUST	10,000	500	20
FEAR	10,000	500	20
HAPPINESS	10,000	500	20
SADNESS	10,000	500	20
SURPRISE	10,000	500	20
NEUTRAL	-	0(500)	0(20)
Total	60,000	3,000(3,500)	120(140)

Table 6. Hyper-parameters of CNN training.

Hyper-parameter name	Value		Description
	Big data	Good data	
test_iter	1,000		Number of times to calculate the loss by extracting the batch with one training. Since the batch number is 16, a loss is calculated for 16 × 1000 = 16000 images. Update the parameter of the convolution layers based on loss.
base_lr	0.0001		(Initial) learning rate. Indicator for varying parameters. The training rate may be changed in the middle according to lr_policy.
lr_policy	"step"		Whether decrease the learning rate when the number of training increases. In the case of "step", the learning rate for the number of learning iter is calculated by Eq. (1).
gamma	0.1		The value used in calculation of learning rate. (See description of lr_policy)
stepsize	100,000		The value used in calculation of learning rate. (See description of lr_policy)
max_iter	500,000	250,000	Number of times to conduct training. It is also called *iteration*.

$$lr\_policy = base\_lr * gamma^{(floor(iter/stepsize))} \quad (1)$$

In contrast, the reason of the high correct answer rate of HAPPINESS and SURPRISE is considered to be obvious difference in the shape of the face with other emotions. In order to solve the above problems and improve the accuracies, it is considered that the following work is necessary to implement.

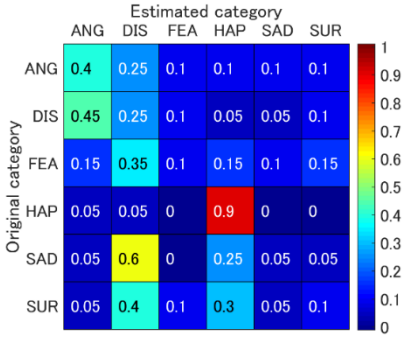
- Increase the number of good data.
- Check the good data and delete the facial emotion images with incorrect emotion tags.

Verification of facial emotion reconstruction using StarGAN

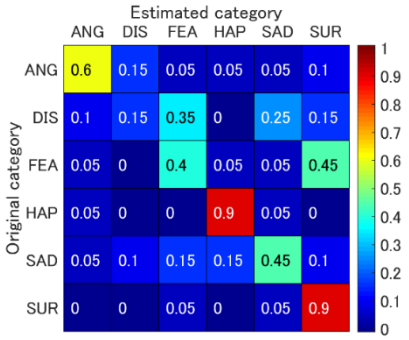
To confirm the usefulness of the good data obtained in this research, we also used the good data for the deep learning-based facial image reconstruction. In this study, we applied the training data of good data to StarGAN [8]. StarGAN is an approach for converting images to multiple domains using only a single model. Implementation by the authors has published in (https://github.com/yunjey/StarGAN).

In this research, we trained on the good data using RaFD training parameters in authors' implementation (input image transformed to 128×128). Figure 7 shows the result of reconstructing the facial emotion of CelebA[15].

From Figure 7, we can reconstruct to the desired emotion by using StarGAN trained with the good data. However, since the image quality of the reconstruction is sometime not satisfied. It is often hard to understand the change in some facial parts. It is necessary to increase the number of the good data for improving the accuracy in the facial emotion reconstruction.



(a) Big data



(b) Good data

Figure 5. Results of six-category classification by networks trained with big data and good data.

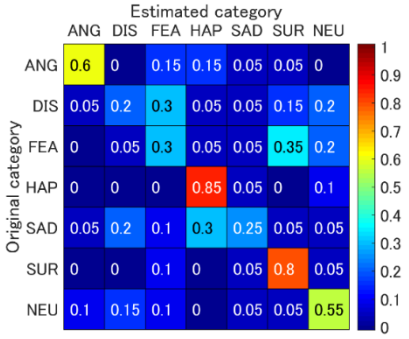


Figure 6. Result of seven-category classification by network trained with good data.

Figure 6. The result of reconstructing the facial emotion of CelebA. The leftmost row is the original images. The other rows are the results of reconstruction. From the left to right, show the result of construction to ANGER, DISGUST, FEAR, HAPPINESS, NEUTRAL, SADNESS and SURPRISE.

Conclusions

In this study, we collected approximately 380,000 facial region images tagged with emotion terms as "big data" from Twitter. We used the emotion terms of 6 categories: ANGER, DISGUST, FEAR, HAPPINESS, SADNESS, and SURPRISE. Next, we got about 5,500 facial region images tagged with seven kinds of emotions including NEUTRAL as "good data" through the subjective experiment. We also confirmed that it was possible to create a classification model with approximately double the accuracy rate, even though the number of the training images of the good data was one-twentieth the number of the big data. However, the accuracy rate is not satisfied (51%). It will be necessary to train again after increasing the number of the good data. Additionally, we confirmed the usefulness of the good data as training data for the facial emotion reconstruction using StarGAN.

As future works, we will increase and refine the good data and verify whether we can achieve deep learning-based facial image processing with higher accuracy. We also consider the application of the facial recognition with multiple emotions from a single expression image and the reconstruction of a facial image with multiple (complex) emotions. In terms of the acquisition of good data, we will consider facial image collections with more complicated emotions beyond the seven emotion categories used in this study.

References

[1] SoftBank Robotics. Softbank Robotics — Humanoid robotics & programmable robots. Retrieved Nov 04, 2018 from https://www.softbankrobotics.com/emea/en

[2] A. Kalegina, G. Schroeder, A. Allchin, K. Berlin, and M. Cakmak. "Characterizing the Design Space of Rendered Robot Faces," in the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18), ACM, New York, NY, USA, pp. 96–104, 2018.

[3] H. Nomiya and T. Hochin, "Facial Expression Recognition using Feature Extraction based on Estimation of Useful Facial Features," Jour. Japan Society for Fuzzy Theory and Intelligent Informatics, vol. 23, issue 2, pp. 170-185, 2011.

[4] A. M. Bukar and H. Ugail. "Facial Age Synthesis Using Sparse Partial Least Squares (The Case of Ben Needham)," Jour. Forensic Sci. vol. 62, issue 5, pp. 1205-1212, 2017.

[5] S. Alizadeh and A. Fazel, "Convolutional Neural Networksfor Facial Expression Recognition," CoRR abs/1704.06756, 2017.

[6] J. Liao, Y. Yao, L. Yuan, G. Hua, and S, B, Kang, "Visual Attribute Transfer through Deep Image Analogy," CoRR abs/1705.01088, 2017.

[7] J. M. Girard and D. McDuff, "Historical Heterogeneity Predicts Smiling: Evidence from Large-Scale Observational Analyses," in Automatic Face and Gesture Recognition (FG), 2017 12th IEEE International Conference on. IEEE, Washington, DC, USA, 2017.

[8] Y. Choi, M.-J. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8789-8797, 2018.

[9] P. Ekman and W. V. Friesen, Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. Prentice Hall, 1975.

[10] TwitterIR, Q1 2017 Letter to Shareholders, Retrieved Nov 05, 2018 from https://investor.twitterinc.com/static-files/d650c15d-5774-4914-95a2-a1562a8e1fd6

[11] Twitter Developers, Twitter Developer Platform, Retrieved Nov 05, 2018 from https://developer.twitter.com/

[12] OpenCV, Retrieved Nov 05, 2018 from http://opencv.jp/

[13] Caffe, Retrieved Nov 05, 2018 from http://caffe.berkeleyvision.org/

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," in the 22Nd ACM International Conference on Multimedia (MM '14). ACM, New York, NY, USA, pp. 675–678, 2014.

[15] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild", in International Conference on Computer Vision (ICCV), 2015.

Author Biography

Tomoyuki Takanashi received his Bachelor of Engineering from Chiba University in 2018. He is currently a master's program student in Graduate School of Science and Engineering, Chiba University. His work has focused on the deep learning, material appearance and subjective quality.

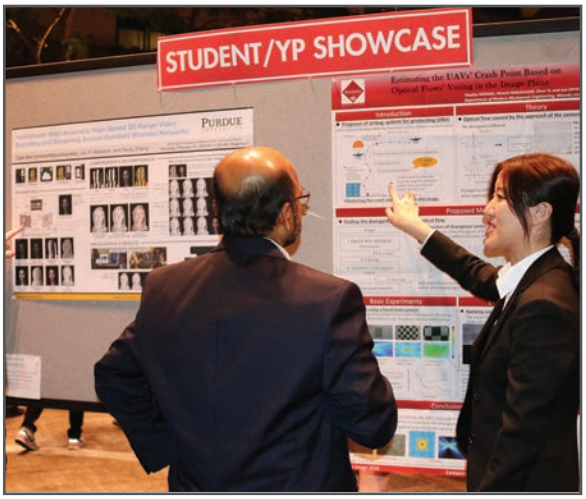
JOIN US AT THE NEXT EI!

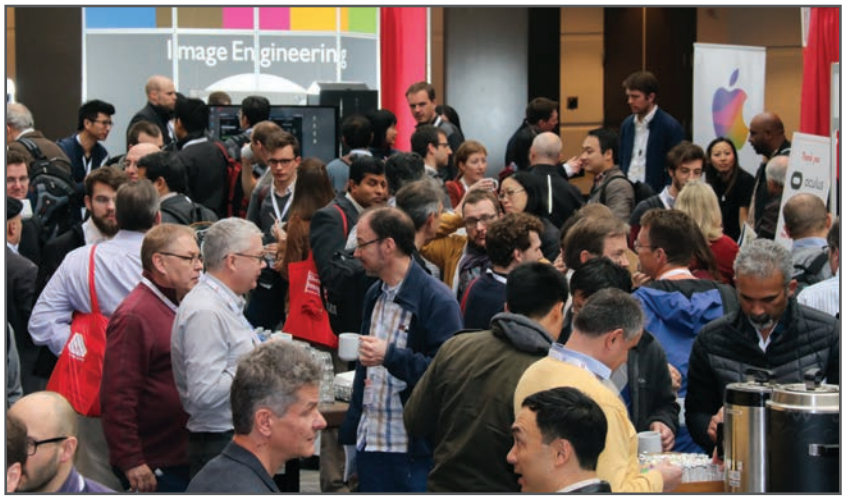
IS&T International Symposium on

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