https://doi.org/10.2352/ISSN.2470-1173.2019.10.IQSP-308

Subjective Evaluations on Perceptual Image Brightness in High Dynamic Range Television

Yoshitaka Ikeda, Yuichi Kusakabe; NHK (Japan Broadcasting Corporation); Tokyo, Japan

Abstract

Viewers of high dynamic range television (HDR, HDR-TV) expect a comfortable viewing experience with significantly brighter highlights and improved details of darker areas on a brighter display. However, extremely bright images on a HDR display are potentially undesirable and lead to an uncomfortable viewing experience. To avoid the issues, we require specific production guidelines for subjective brightness to ensure brightness consistency between and within programs. To create such production guidelines, it is necessary to develop an objective metric for subjective brightness in HDR-TVs. A previous study reports that the subjective brightness is proportional to the average of displayed pixel luminance levels. However, other parameters can affect the subjective brightness. Therefore, we conducted a subjective evaluation test by using specific test images to identify the factors that affect the perceived overall brightness of HDR images. Our results indicated that positions and distributions of displayed pixel luminance levels on video affect brightness in addition to the average of displayed pixel luminance levels. The study is expected to contribute to the development of an objective metric for subjective brightness.

Introduction

In Japan, 4K and 8K ultrahigh-definition television (UHDTV) broadcasting via satellite commenced on December 1, 2018. Broadcasts on UHDTV feature ultrahigh-definition images and also wide color gamut and high dynamic range (HDR). The specifications of UHDTV are standardized by the International Telecommunication Union – Radio communication Sector (ITU-R) [1] [2]. Recommendation ITU-R BT.2100 (BT.2100) specifies HDR television (HDR-TV) image parameters for use in production and international program exchange by using the Perceptual Quantization (PQ) and Hybrid Log-Gamma (HLG) methods [2]. HDR-TV systems extend the range of luminance, and this provides content producers with additional freedom to make content creative with respect to brightness. Viewers of HDR-TV expect a comfortable viewing experience with significantly brighter highlights and improved details of darker areas on a brighter display. However, extremely bright images on a HDR display can be undesirable and lead to an uncomfortable viewing experience. Greater brightness jumps can occur between and within programs irrespective of the content producer's intent. To avoid the aforementioned issues, we require specific production guidelines for subjective brightness to ensure brightness consistency between and within programs in a manner similar to audio loudness. To create such production guidelines, it is necessary to develop an objective metric for subjective brightness in HDR-TVs.

A previous study [3] reported that the subjective brightness is proportional to the average of displayed pixel luminance levels (ALL). However, there can exist images that differ in subjective brightness although they exhibit the same ALL. Additionally, other parameters can affect subjective brightness.

The study involves two objectives: one is to investigate if the ALL alone is sufficient to represent brightness while viewing HDR images, and the other is to estimate other possible factors that determine brightness in addition to the ALL.

Background

Operational Practice in Television Production

In standard dynamic range (SDR) television, an operational practice in television production is not officially standardized. As described in [3], conventional tone mapping that maps specific scene luminance levels to appropriate signal levels is widely used. The range of luminance is limited and the peak luminance level is not extremely high in SDR television, and thus there is no need to specify an operational practice or a metric for brightness.

In HDR-TV, the range of luminance is significantly wider, and thus an operational practice in HDR-TV production is needed. Report ITU-R BT.2408 [4] specifies the operational practice in HDR-TV production and provides initial guidance to help ensure the optimal and consistent use of HDR via the PQ and HLG methods. With respect to HLG, a reference white level of 75% is recommended to ensure sufficient headroom for specular highlights and maintain some consistency of brightness between programs. However, to avoid unexpected brightness jumps between and within programs, we require specific production guidelines for subjective brightness. To create such production guidelines, it is necessary to develop an objective metric for subjective brightness in HDR-TVs.

Brightness Perception

Stevens proposed a power law correlation between the luminance and brightness [5] [6]. The results of his tests indicated that the value of the power law varies based on the apparent sizes of stimuli and visual adaptation states and that a power of 0.33 of relative luminance is proportional to the perceived brightness. Bauer indicated that Stevens' power law can be extended to estimate the perceived brightness of a group of several patches [7]. In those studies, the stimuli were small and uniform. Conversely, an objective metric which we aim to develop in the present study can be applied to various complex images (i.e., natural images), patterns that exhibit particular dark areas or highlights, and any other content on television.

To determine the subjective brightness of overall images, Zipa et al. proposed an algorithm to estimate the subjective brightness of overall images wherein pixel luminance levels differed [8]. They considered both local and global brightness in the algorithm with reference to the CIECAM02 [9]. They insisted that the algorithm was more suitable to represent brightness than simple equations such as the ALL. However, it was extremely complicated for application to real-time instruments in television production.

To determine a real-time objective metric for overall brightness of HDR images, Noland et al. conducted a subjective evaluation test [3]. They proposed several possible metrics and compared them with each other. The results indicated that the ALL is the most

accurate and simplest metric for real-time instruments in HDR-TV production. Chapiro et. al. reported that brightness is related to the ALL and also the physical size and distance from the screen [10]. However, these studies did not determine whether there are significant differences in brightness among several sets of test images that exhibited the same ALL. This implies that there can be other possible factors that determine brightness in addition to the ALL. Therefore, in the present study, we conducted a subjective evaluation test by using specific test images to identify the factors that affect the perceived overall brightness of HDR images.

Prior to our subjective evaluation test, we estimated the factors that determine brightness. First, the ALL was considered as the primary factor. In addition to ALL, three factors were considered as possible factors. The first factor corresponded to the position of displayed pixel luminance because the displayed pixel luminance closer to the center more significantly affects brightness. The second factor corresponded to the distribution of displayed pixel luminance wherein high luminance pixels close to each other can exhibit a more significant influence. The last factor corresponded to the contrast, and this implies that a high contrast image can be perceived as a brighter image. We prepared the test images such that the factors could be fairly analyzed.

Experiment

Table 1 shows an overview of experimental conditions. Fifteen video researchers participated in the experiment. A test image and a full-screen gray pattern were alternately displayed. Each subject was asked whether a gray pattern is "brighter" or "darker" or "equal to" when compared with a test image. An experimental setup is presented in the following section. All images used in our test are shown in the "Images" section. The experiment was conducted via a method of limits [11] and is described in detail in the "Experimental Procedure" section.

Experimental Setup

Figure 1 shows a diagram of the experimental setup. A 29.5-inch 4K HDR OLED mastering monitor (Sony BVM-X300) with a peak luminance of 1,000 cd/m^2 was used to display all the images.

Table 1: Experimental conditions

Experimental Procedure	Method of limits
Test images	Twenty 4K still images (14 patterns and six natural images, including two images for instructions and training)
Gray patterns	- 27 steps - Flat patterns in the range 2.9–154.9 cd/m^2 - Equal intervals in the lightness scale
Viewing condition	- Dim surround - 4K OLED display (peak luminance: 1,000 cd/m^2) - 1.5 times the picture height
Subjects	15 video researchers

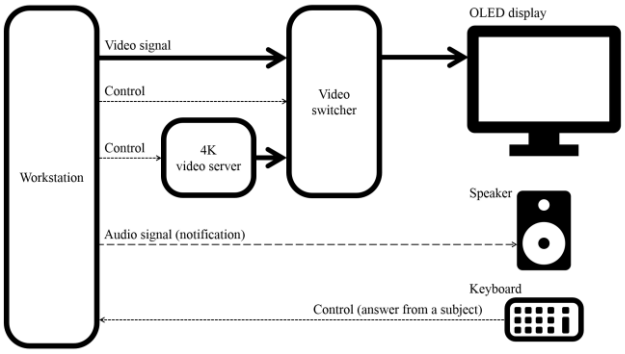


Figure 1. Diagram of experimental setup

The HLG electro-optical transfer function was set to the OLED display. A workstation (Dell Precision T7500) loaded with a 4K playback PCIe card (Blackmagic Design DeckLink 4K Pro) played and switched a gray pattern based on the subject's answers that were entered via a keyboard. Additionally, the workstation assumed control of a video switcher (Blackmagic Design ATEM 2 M/E Broadcast Studio 4K) via an ethernet cable and a 4K video server (Astrodesign HR-7512-C) via a 9-pin D-sub cable. The 4K video server accepted the commands as RS-422 protocols converted from RS-232 protocols. Both video signals from the workstation and 4K video server were input to the OLED display through the video switcher. A speaker (Fostex NetCIRA ES6300) was positioned beside a subject to emit sounds of the notification signals to inform the subject that the workstation accurately accepted each operation via the keyboard.

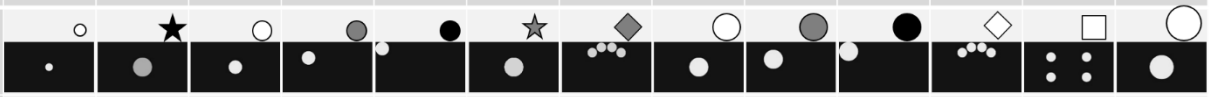
Figure 2 shows the test room setup. The test room was carefully arranged in compliance with the reference viewing environments specified in BT.2100 [2]. The room was dimly surrounded, and all the walls were covered with curtains. The area surrounding a display was gray, and the other area was black. The OLED display was placed on a table covered by a black cloth. The viewing distance was set as 1.5 times the picture height. Two adjustable LED lights (Flolight MicroBeam 512) illuminated the gray curtain behind the OLED display. The light reflected off the wall measured D65 white at approximately 5 cd/m^2. The LED lights were positioned under the table such that the direct light did not enter a participant's eyes.

Images

Table 2 shows twenty test images and their parameters with the markers as prepared for use in scatter plots in the next section.

Figure 2. Test room setup

Table 2: Test Images

No.	1	2	3	4	5	6	7	8	9	10	11	12	13
Image													
ALL [cd/m^2]	10.0	←	31.6	←	←	←	←	63.1	←	←	←	←	100.0
Position			center	mid	corner			center	mid	corner			
Distribution									high		mid	low	
Contrast	high	low	high			low							
Maximum Luminance [cd/m^2]	1000	158	1000	←	←	501	←	1000	←	←	←	←	←
Minimum Luminance [cd/m^2]	0.10	←	←	←	←	←	←	←	←	←	←	←	←

No.	14	15	16	17	18	19	20
Image							
ALL [cd/m^2]	10.0	31.6	←	63.1	100.0	49.8	63.1
Maximum Luminance [cd/m^2]	754	750	1000	967	1000	←	←
Minimum Luminance [cd/m^2]	0.10	←	←	←	←	←	←

Images Nos.19 and 20 were prepared for instructions and training of the experimental procedure with no marker indicated at these two images in the Table.

Images Nos.1–13 and 20 corresponded to patterns created by us, and exhibit one or four white patches on a black background. There were four groups of ALLs, namely 10.0, 31.6, 63.1, and 100.0 cd/m^2. All the patterns exhibited a black background with a luminance of 0.1 cd/m^2. The differences in ALLs between the patterns were derived from the luminance and sizes of the white patches. In each of the two groups with ALLs of 31.6 and 63.1 cd/m^2, there were three images with a patch each at different positions (Nos.3–5 with the ALL of 31.6 cd/m^2 and Nos.8–10 with the ALL of 63.1 cd/m^2). They were prepared for the investigation as to whether the position of the displayed pixel luminance affected brightness. In the group with the ALL of 63.1 cd/m^2, there were three images in which the distributions of displayed pixel luminance differed (Nos.9, 11, and 12). They were prepared to verify whether the distribution of the displayed pixel luminance affected brightness. In each of the groups with ALLs of 10.0 and 31.6 cd/m^2, there were two images in which the contrasts differed (Nos.1 and 2 with the ALL of 10.0 cd/m^2, Nos.3 and 6 with the ALL of 31.6 cd/m^2). Both the positions and sizes of the patches included in the images Nos.2 and 6 were identical to that of the image No.8. Image No.20 was identical to No.12.

Images Nos.14–19 corresponded to natural images shot by us via a 4K HDR camera (SONY HDC-4300) with a HLG opto-electronic transfer function. Among images Nos.14–18, there were four groups with different ALLs that were identical to the groups of the patterns described above. Image No.19 exhibited the ALL of 49.8 cd/m^2.

Specifically, 27 gray patterns were prepared for the measurement of brightness of the test images. These were flat patterns that ranged from 2.9 to 154.9 cd/m^2. The steps were set at equal intervals in the lightness scale.

Procedure

The tests for each subject were separated by twenty sessions. Each session was conducted via a method of limits [11] and consisted of a pair of "up series" and "down series." In an up series, a gray pattern with the lowest luminance was initially displayed, and a gray pattern became 1-step brighter at each time a subject answered via a keyboard. The subject was asked to push an enter key on the keyboard to commence the first trial in each series. Subsequently, a gray pattern and a test pattern were alternately displayed every 3 s until the subject answered via the keyboard as to whether the gray pattern was "brighter," "darker" or "equal to" when compared with the test pattern in each trial of each series. There was no time limit to answer. When an answer was given, the gray pattern switched 1-step-up to the next luminance such that the subject could move on to the next trial. In the next trial, the subject answered by following the same procedure outlined above. The up series continued until the subject answered "brighter" for the first time in the series. A down series was conducted in the same manner

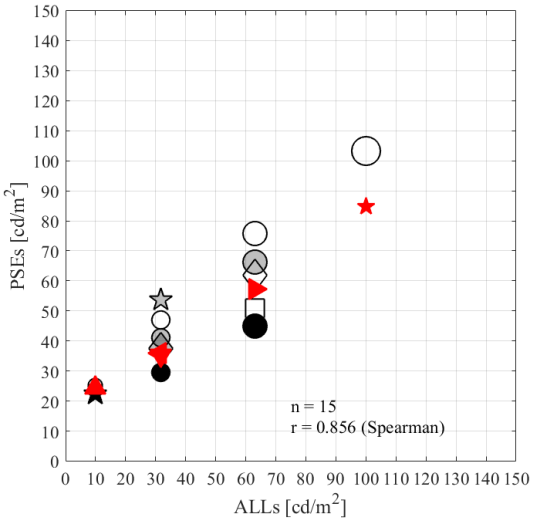


Figure 3a. Relationship between the ALLs and average PSEs

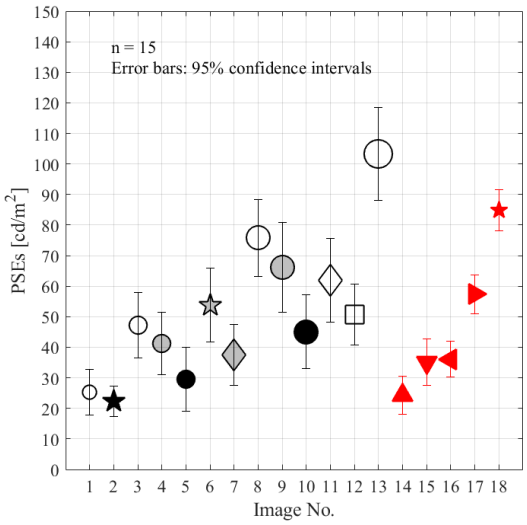


Figure 3b. Average PSEs per each test images

albeit in the opposite direction as an up series. In a down series, a gray pattern with the highest luminance was initially displayed, and this became 1-step darker at each time a subject answered. The series continued until the subject answered "darker" for the first time in it.

Each of the twenty test images were assigned in each session. Test images Nos.19 and 20 were routinely assigned for the first two sessions for each subject. These sessions were set as instructions and training for the procedure. Each of the test images Nos.1–18 was assigned to each of the third and following sessions at random. In each session, the order of up/down series was also set randomly.

The point of subjective equality (PSE) in brightness was taken as the average of 4 transition points from each subject's answers in a pair of series in each session, namely the highest and the lowest luminance levels of gray patterns within the range of "equal to." The PSEs were used as the perceived overall brightness of each test image to determine an objective metric for subjective brightness in HDR-TVs.

Results and Discussion

Figure 3a shows the relationship between the ALLs of the eighteen test images and average PSEs of all the subjects. Each marker corresponds to each test image shown in Table 2, respectively. The results indicated correlation between the ALLs and PSEs with a Spearman's rank correlation coefficient of 0.856. However, PSEs of several images that exhibit the ALLs of 31.6 and 63.1 cd/m^2 appeared as spread in a wide range although they exhibit the same ALLs. To ensure that each marker is more easily viewed, the average PSEs per each test images are shown in Figure 3b. The error bars exhibited 95% confidence intervals. The plot indicated that the three possible factors as described in the "Brightness Perception" section can influence brightness: the PSEs exhibited differences among Nos.3–5 with the same ALL of 31.6 cd/m^2 and Nos.8–10 with the same ALL of 63.1 cd/m^2 (caused by the position), among Nos.9, 11, and 12 with the same ALL of 63.1 cd/m^2 (caused by the distribution), between Nos.1 and 2 with the ALL of 10.0 cd/m^2, and between Nos.3 and 6 with the ALL of 31.6 cd/m^2 (caused by the contrast).

Hence, the results confirmed via a paired t-test whether the differences in brightness were statistically significant. The null hypothesis stated that brightness is identical for two images with the same ALL. First, with respect to a pair of the images that exhibit the same ALLs but exhibit a white patch at different positions, the null hypothesis was rejected (Nos.3 and 5, $t(14) = 3.318$, $p < 0.01$; Nos.4 and 5, $t(14) = 2.317$, $p < 0.05$; Nos.8 and 10, $t(14) = 8.751$, $p < 0.01$; Nos.9 and 10, $t(14) = 4.471$, $p < 0.01$). This implied that the differences in brightness caused by position are statistically significant. Second, with respect to a pair of the images that exhibit the same ALLs albeit exhibiting white patches distributed differently, the null hypothesis was also rejected (Nos.9 and 11, $t(14) = 2.586$, $p < 0.05$; Nos.9 and 12, $t(14) = 2.705$, $p < 0.05$). This implied that the differences in brightness caused by distribution are statistically significant. Finally, with respect to any pair of the images that exhibit the same ALLs albeit exhibiting different contrast, the null hypothesis was not rejected at $p = 0.05$ level. The results of the paired t-tests are summarized as follows: the differences in brightness due to the positions and distributions are statistically significant, and this indicates that the ALL alone is not sufficient to represent brightness while viewing HDR images.

To determine a more accurate metric to represent brightness, we implemented the equation "corrected ALL" that considers the effect of position shown as below:

$$\text{Corrected ALL} = \frac{1}{\theta_{\text{mean}}} \cdot \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(L_{i,j} \cdot \cos^3 \theta_{i,j} \right)$$

where i and j denote pixel indices with $i \in$ 0:M–1 and $j \in$ 0:N–1. With respect to the test images used in our test, $M = 2160$ and $N = 3840$. Specifically, L denotes the displayed pixel luminance level at pixel (i, j), and θ denotes the angle subtended at the eye between pixel (i, j) and center of the screen as shown in Figure 4. The value of θ is dependent on the viewing distance. The value of the third power of cosθ was derived from two physical factors, namely the solid angle and tilt angle. θ_{mean} denotes the mean value of $\cos^3\theta$ calculated in all the pixels and is defined as the normalizing factor. In our test, $\theta_{\text{mean}} = 0.8197$.

Figure 5 shows the relationship between the corrected ALLs of the eighteen test images and the average PSEs of all the subjects. The results indicated correlation between the corrected ALLs and PSEs with a Spearman's rank correlation coefficient of 0.969, which evidently exceeds 0.856 as calculated in Figure 3a. To evaluate the significance of the differences between two values above, we

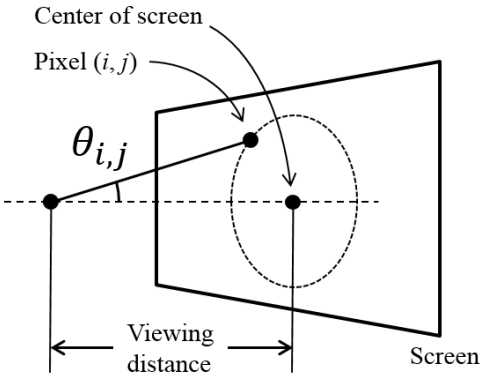


Figure 4. Definition of angle θ per each pixel

conducted a test for independent correlation coefficients. The results revealed that the difference between the two values was statistically significant at $p = 0.01$ level. This clearly indicated that the correction by the weighting factor of the $\cos^3\theta$ was effective.

Additionally, we evaluated the differences caused by the distribution. As mentioned in the previous section, images Nos.9, 11, and 12 exhibit different distributions of displayed pixel luminance although they exhibit the same ALL of 63.1 cd/m^2. The data points of the three images as shown in Figure 5 indicate they also have the same corrected ALL of approximately 70 cd/m^2 but exhibit different PSEs. This implied that differences in brightness exist and are caused by distribution among images that exhibit the same corrected ALLs. Therefore, the differences caused by distribution among the three images were compared after calculating the corrected ALLs. We set up the null hypothesis for a paired t-test wherein brightness is identical for two images with the same corrected ALL. As a result, the null hypothesis was rejected and differences in brightness between two pairs of the three images were statistically significant (Nos.9 and 11, $t(14) = 2.586$, $p < 0.05$; Nos.9 and 12, $t(14) = 2.705$, $p < 0.05$). This suggested the need to correct the effect of the distribution in addition to the effect of the position. The definite weighting factor for the distribution will be examined in a future study.

Conclusion

To develop an objective metric for subjective brightness in HDR-TVs, we conducted a subjective evaluation test for brightness while viewing HDR images. The results indicated that there are differences in brightness even if the images exhibit the same ALL. Therefore, the ALL alone is not a suitable metric to represent brightness. Additionally, the result suggested that the corrections by position and distribution of displayed pixel luminance are effective. The equation for "Corrected ALL", which includes a weighting factor of $\cos^3\theta$ to compensate the effect of position, could be a better metric to represent brightness than the ALL. A future study will involve determining a definite weighting factor for distribution.

References

[1] ITU-R, "Parameter values for ultra-high definition television systems for production and international programme exchange," Recommendation ITU-R BT.2020-2, October 2015.

[2] ITU-R, "Image parameter values for high dynamic range television for use in production and international programme exchange," Recommendation ITU-R BT.2100-2, July 2018.

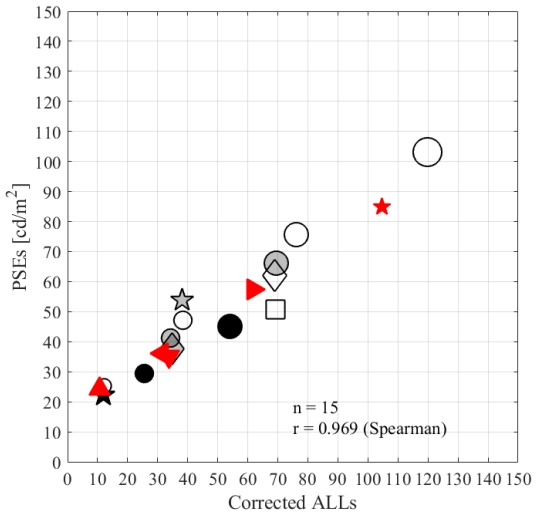


Figure 5. Relationship between the corrected ALLs and PSEs

[3] K. C. Noland, M. Pindoria and A. Cotton, "Modelling Brightness Perception for High Dynamic Range Television," in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, pp. 1–6, May–June 2017.

[4] ITU-R, "Operational practices in HDR television production," Report ITU-R BT.2408-0, October 2017.

[5] S. S. Stevens, "On the Psychophysical Law," The Psychological Review, vol. 64, no. 3, pp. 153–181, 1957.

[6] S. S. Stevens, "To Honour Fechner and Repeal His Law," Science, vol. 133, no. 3446, pp. 80–86, 1961.

[7] B. Bauer, "Does Stevens' Power Law for Brightness Extend to Perceptual Brightness Averaging?" The Psychological Record, vol. 59, pp. 171–186, 2009.

[8] K. Zipa and A. Ignatenko, "Estimation of Object's Integral Brightness," in 2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, pp. 359–366, September 2015.

[9] CIE. Publ, 159-2004, "A colour appearance model for colour management systems: CIECAM02," CIE Central Bureau: Vienna, 2004.

[10] A. Chapiro, T. Kunkel, R. Atkins and S. Daly, "Influence of Screen Size and Field of View on Perceived Brightness," ACM Transactions on Applied Perception, vol. 15, no. 3, pp. 1–13, August 2018.

[11] T. Kuroda and E. Hasuo, "The Very First Step to Start Psychophysical Experiments," Acoustical Science and Technology, vol. 35, no. 1, pp. 1–9, 2014.

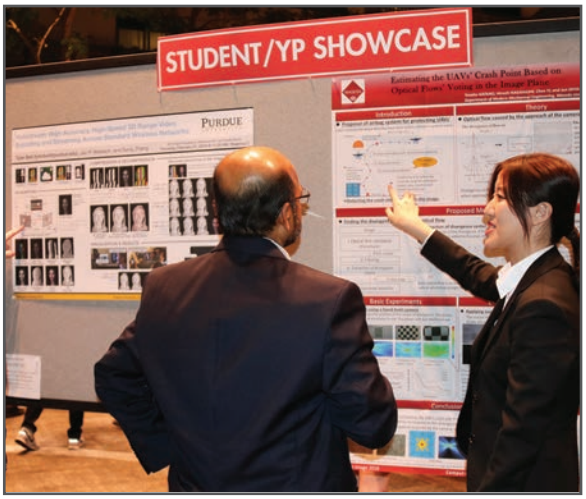
JOIN US AT THE NEXT EI!

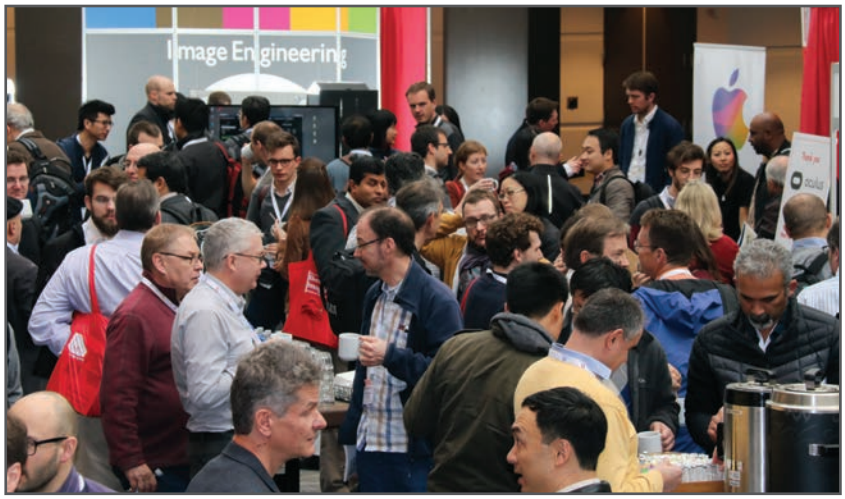
IS&T International Symposium on

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