https://doi.org/10.2352/ISSN.2470-1173.2020.9.IQSP-167


# Analyzing the performance of autoencoder-based objective quality metrics on audio-visual content

***Helard Becerra Martinez*, Mylène C.Q. Farias+ and Andrew Hines*;***
** School of Computer Science, University College Dublin, Dublin, Ireland;*
*+ Department of Electrical Engineering, University of Brasília, Brasília, Brazil*

## Abstract

*The development of audio-visual quality models faces a number of challenges, including the integration of audio and video sensory channels and the modeling of their interaction characteristics. Commonly, objective quality metrics estimate the quality of a single component (audio or video) of the content. Machine learning techniques, such as autoencoders, offer as a very promising alternative to develop objective assessment models. This paper studies the performance of a group of autoencoder-based objective quality metrics on a diverse set of audio-visual content. To perform this test, we use a large dataset of audio-visual content (The UnB-AV database), which contains degradations in both audio and video components. The database has accompanying subjective scores collected on three separate subjective experiments. We compare our autoencoder-based methods, which take into account both audio and video components (multi-modal), against several objective (single-modal) audio and video quality metrics. The main goal of this work is to verify the gain or loss in performance of these single-modal metrics, when tested on audio-visual sequences.*

## Introduction

The popularity of multimedia services and the massive consumption of multimedia content, through wired and wireless internet-based networks, has increased the interest in the area of user quality of experience. More specifically, given that the success and popularity of multimedia services is correlated with the quality of the content experienced by the end-user, over the last decades researchers have proposed several objective quality metrics that automatically estimate the quality of the signal at the user end [1]. Among the available metrics, models that are based on the visual and auditory human systems are very appealing since they produce estimates that are better correlated with actual human responses collected on subjective experiments (i.e. subjective quality scores). However, to our knowledge, existing contributions are single modality (single-modal) quality assessment methodologies, i.e., audio-only or video-only objective quality metrics. So far, very few proposals have tackled the multi-modal problem, even for the simpler case of audio-visual quality. It is worth pointing out that the development of audio-visual objective quality metrics face a number of challenges, like for example the integration of the audio and video sensory channels and its corresponding perceptual modeling. Some proposals use a parametric approach, which consists of using encoding and networking parameters to predict the audio-visual quality [2]. However, this type of solution is restrictive because of its dependency on the system parameters. One way of dealing with the audio and video integration problem is to use machine-learning approaches. In particular, autoencoder-based approaches can be used to train a complex function that integrates a set of descriptive audio and video features and, then, creates a set of new features. This type of approach is an interesting and promising way of facing the multi-modal quality assessment problem [3]. Very few studies have explored the performance of single quality models on appropriate multi-modal (e.g., audio-visual) material, containing degradations in the different components. Therefore, there is a need for performance studies that can determine the level of accuracy of single and multi-modal quality metrics on datasets containing audio and visual distortions.

In this paper, we intend to evaluate the performance of quality assessment methodologies on large audio-visual datasets. To this end, models using an autoencoder approach are tested against a number of well know audio-only and video-only (single modality) objective quality metrics. The audio-visual datasets contain common audio and visual distortions and a large variety of audio and visual content. In summary, this study has the goal of: 1) verifying the performance of multi-modal metrics on audio-visual material impaired with both audio and video distortions, and 2) verifying the performance and possible gains or losses of single-modal objective metrics on audio-visual material impaired with audio and video distortions.

The structure of this paper is organized as follows. First, the audio-visual dataset used for this study is presented. Then, the objective metrics from the literature are listed along with the autoencoder-based metrics used in this study. Next, the objective results gathered for all the objective metrics are presented and commented. Finally, some conclusions of the study are presented.

## Audio-visual Material

To study the performance of objective quality models, we need a large set of audio-visual content, with their accompanying subjective responses. This content must reflect the scope of common multimedia applications, that is, they need to consider: common types of multimedia components (audio and video), common multimedia scenarios and the resulting types of degradation, and a diverse source content (e.g., video conferencing, movies, sports transmissions, documentaries, etc.).

With this goal, we use the UnB-AV database [4], which is composed of content that was subjectively rated in three separate experiments. This database contains a variety of audio-visual content sources, ranging from various video genres like movie trailers, sports, TV commercials, interviews, etc. Figure 1 shows sample video frames of the database source content. For all three experiments, groups of human observers rated the audio-visual

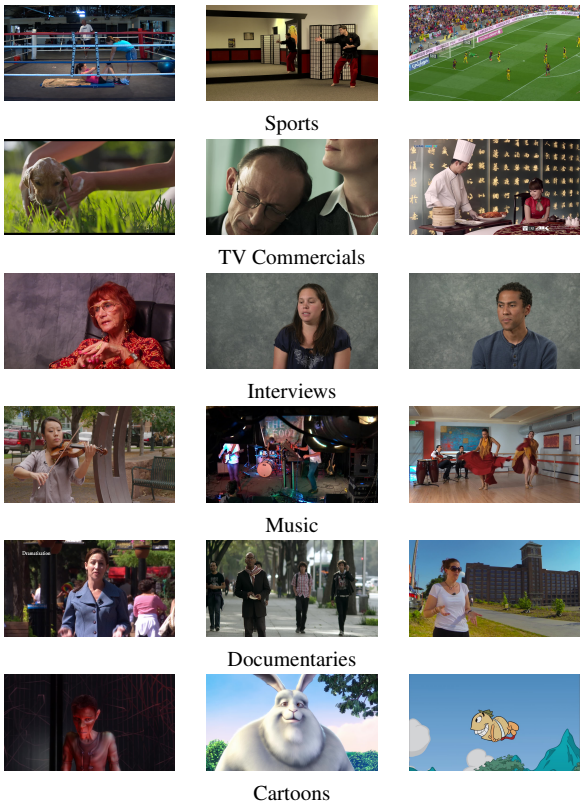



Figure 1: Sample frames of the videos from the UnB-AV Database. The database include different genres contents like: Sports, TV Commercials, Interviews, Music, Documentaries, and Cartoons.

quality of the corresponding set of audio-visual sequences.

The UnB-AV database has sets of visual and audio distortions, commonly found in multimedia consumption systems. Visual distortions include video coding, packet loss, and frame freezing, while audio distortions include background noise, clipping, echo, and chop. For the first experiment, only the video component is degraded, meanwhile the audio component does not have any type of degradation. In the second experiment, the audio component was degraded, while the video component remained untouched. Finally, in the third subjective experiment, both audio and video components were degraded. The test conditions for all three experiments are detailed in Tables 1, 2 and 3. Details about the set-up of these experiments can be found in a previous work [5].

Table 1: Test conditions for visual degradations from Experiment 1.

| HRC | Codec | Bitrate (kb/s) | Packet Loss PLR | Freezing # Pauses | Freezing P. Length | Freezing P. Position |
|---|---|---|---|---|---|---|
| $HRC1_{E1}$ | H.264 | 500 | 10% | - | - | - |
| $HRC2_{E1}$ | H.265 | 400 | 8% | - | - | - |
| $HRC3_{E1}$ | H.264 | 2000 | 5% | - | - | - |
| $HRC4_{E1}$ | H.265 | 1000 | 3% | - | - | - |
| $HRC5_{E1}$ | H.265 | 8000 | 1% | - | - | - |
| $HRC6_{E1}$ | H.265 | 200 | - | 3 | 3, 3, 2 | 1, 2, 3 |
| $HRC7_{E1}$ | H.264 | 800 | - | 3 | 2, 2, 3 | 1, 2, 3 |
| $HRC8_{E1}$ | H.265 | 1000 | - | 2 | 2, 2 | 2, 3 |
| $HRC9_{E1}$ | H.264 | 2000 | - | 2 | 1, 3 | 1, 3 |
| $HRC10_{E1}$ | H.264 | 16000 | - | 1 | 2 | 2 |
| $ANC1_{E1}$ | H.264 | 64000 | - | - | - | - |
| $ANC2_{E1}$ | H.265 | 32000 | - | - | - | - |

Table 2: Test conditions for audio degradations from Experiment 2.

| BG Noise | Noise | SNR (dB) | |
|---|---|---|---|
| $HRC1_{E2}$ | car | 15 | |
| $HRC2_{E2}$ | babble | 10 | |
| $HRC3_{E2}$ | office | 10 | |
| $HRC4_{E2}$ | road | 5 | |
| $ANC1_{E2}$ | - | - | |
| **Chop** | **Period (s)** | **Rate (chops/s)** | **Mode** |
| $HRC5_{E2}$ | 0.02 | 1 | previous |
| $HRC6_{E2}$ | 0.02 | 2 | zeros |
| $HRC7_{E2}$ | 0.04 | 2 | previous |
| $HRC8_{E2}$ | 0.02 | 5 | zeros |
| $ANC2_{E2}$ | - | - | - |
| **Clipping** | | **Multiplier** | |
| $HRC9_{E2}$ | | 11 | |
| $HRC10_{E2}$ | | 15 | |
| $HRC11_{E2}$ | | 25 | |
| $HRC12_{E2}$ | | 55 | |
| $ANC3_{E2}$ | | - | |
| **Echo** | **Alpha (%)** | **Delay (ms)** | **Feedback (%)** |
| $HRC13_{E2}$ | 0.5 | 25 | 0 |
| $HRC14_{E2}$ | 0.3 | 100 | 0 |
| $HRC15_{E2}$ | 0.175 | 140 | 0.8 |
| $HRC16_{E2}$ | 0.3 | 180 | 0.8 |
| $ANC4_{E2}$ | - | - | - |

## Objective Metrics

Subjective quality scores from all three experiments are compared against objective scores obtained using a number of well-known Full Reference (FR) and No-Reference (NR) video and audio quality metrics. The FR video quality metrics considered are SSIM [6] and PSNR. The NR video metrics considered are VIIDEO [7], DIIVINE [8], BIQI [9], NIQE [10], and BRISQUE [11]. As for the audio quality metric, the subjective quality scores are compared with the results obtained with a set of FR and NR audio and speech quality metrics from the literature. The FR audio quality metrics considered are VISQOLAudio [12] and PEAQ [13], and the speech metric VISQOL [14]. Finally, the NR speech quality metric P.563 [15] is also considered.

In addition, a set of autoencoder-based objective metrics is considered in our tests: 1) NAVE, for video quality [16], 2) AQUA, for audio quality [17], and 3) NAVIDAD, for audio-visual quality [18]. All three metrics possess the same architecture. They were trained using a two-layer autoencoder plus a classification function, as shown in the block diagram in Figure 2.

The autoencoder-based objective metrics share the same three-layer architecture design presented in Figure 2. At the first stage, a set of descriptive features (audio, video, or both) is extracted from the signal under analysis. At the second stage, the set of features is used as input for a deep net module. This module is formed by two sub-layers: an autoencoder layer and a classification layer. The autoencoder layer takes the extracted set of features and produces a new set of features, which is expected to have a lower dimension and a better description capacity. Next, at the classification layer, the new set of features is mapped into a quality class. At the last stage, results are properly scaled into a <1-5> range, where 1 is considered as the lowest quality score and 5 is considered the highest. A more detailed description of these metrics, that includes feature extraction, training parameters, and testing details, can be found in previous works such as [16, 17, 18]

Table 3: Test conditions for visual and audio degradations from Experiment 3.

| | Audio Component | | | | Video Component | | | |
|---|---|---|---|---|---|---|---|---|
| HRC | Noise Type, SNR (dB) | Chop Period (s), Rate (chop/s), Mode | Clip Multiplier | Echo Alpha (%), Delay (ms), Feedback (%) | Video Codec | Bitrate (kbps) | PacketLoss PLR | Freezing Pauses, Length (s) |
| $HRC1_{E3}$ | car, 15 | - | - | - | H.264 | 16,000 | - | 1, 2 |
| $HRC2_{E3}$ | - | - | 11 | - | H.264 | 16,000 | - | 1, 2 |
| $HRC3_{E3}$ | - | - | 11 | - | H.265 | 8,000 | 0.01 | - |
| $HRC4_{E3}$ | - | 0.02, 2, zeros | - | - | H.265 | 80,00 | 0.01 | - |
| $HRC5_{E3}$ | - | - | - | 0.3, 100, 0 | H.264 | 16,000 | - | 1, 2 |
| $HRC6_{E3}$ | office, 10 | - | - | - | H.264 | 16,000 | - | 1, 2 |
| $HRC7_{E3}$ | - | - | - | 0.3, 100, 0 | H.265 | 8,000 | 0.01 | - |
| $HRC8_{E3}$ | - | - | - | 0.3, 100, 0 | H.264 | 2,000 | 0.05 | - |
| $HRC9_{E3}$ | office, 10 | - | - | - | H.264 | 2,000 | 0.05 | - |
| $HRC10_{E3}$ | office, 10 | - | - | - | H.264 | 800 | - | 3, 7 |
| $HRC11_{E3}$ | - | - | 25 | - | H.264 | 2,000 | 0.05 | - |
| $HRC12_{E3}$ | - | - | 25 | - | H.264 | 800 | - | 3, 7 |
| $HRC13_{E3}$ | - | - | 25 | - | H.265 | 400 | 0.08 | - |
| $HRC14_{E3}$ | - | 0.02, 5, zeros | - | - | H.265 | 400 | 0.08 | - |
| $HRC15_{E3}$ | - | - | - | 0.3, 180, 0.8 | H.264 | 800 | - | 3, 7 |
| $HRC16_{E3}$ | - | - | - | 0.3, 182, 0.8 | H.265 | 400 | 0.08 | - |
| $ANC1_{E3}$ | - | - | - | - | H.264 | 64,000 | - | - |
| $ANC2_{E3}$ | - | - | - | - | H.265 | 32,000 | - | - |
| $ANC3_{E3}$ | - | - | - | - | H.264 | 64,000 | - | - |
| $ANC4_{E3}$ | - | - | - | - | H.265 | 32,000 | - | - |

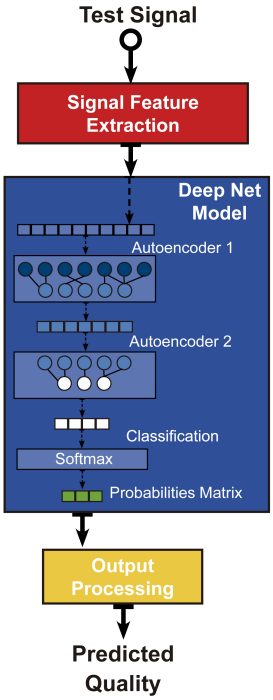
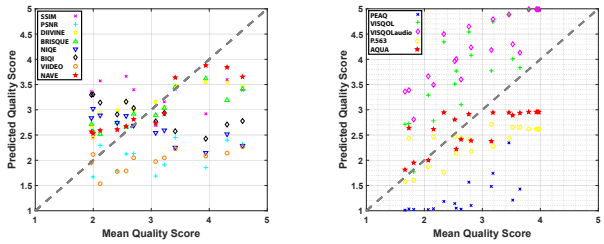



Figure 2: Basic architecture of the Autoencoder-based objective metrics (NAVE, AQUA, NAVIDAD).

## Results

As mentioned earlier, the predicted quality scores from each objective quality metric were compared against the subjective quality scores (Mean Quality Scores MQS) gathered from all three experiments. Figure 3 (a) shows the predicted scores from all video quality metrics versus the collected scores from experiment 1. It can be observed that some visual metrics performed better than others in the audio-visual material. PSNR, SSIM, and BIQI seemed to produce scores that are sparser, while DIIVINE, BRISQUE, and NAVE scores correlated better with the subjective scores.

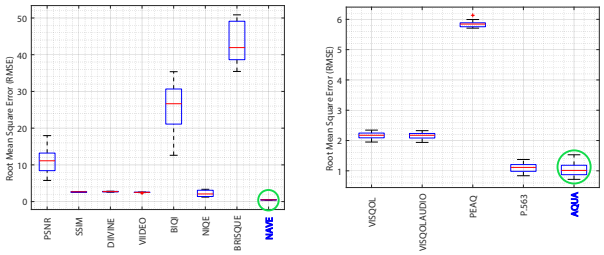
Figure 3 (b) shows the predicted scores obtained using the audio quality metrics versus the subjective scores (MQSs) from experiment 2. From this figure, it can be observed that VISQOL and VISQOLAudio overestimated the quality from sequences, ranking them higher. On the other hand, PEAQ predicted lower quality values for the same audio-visual material. Surprisingly,




(a) (b)

Figure 3: Scatter Plot presenting a Subjective-Objective comparison. (a) Subjective responses from Experiment 1 against a set of visual quality metrics from literature. (b) Subjective responses from Experiment 2 against a set of audio/speech quality metrics.

P563 and AQUA presented a much better correlation with the subjective scores.



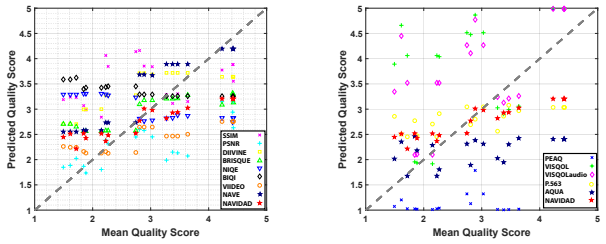

(a) (b)

Figure 4: Bar Plot presenting Root Mean Square Error (RMSE) comparison. (a) Error between subjective responses from Experiment 1 and a set of visual quality metrics from literature. (b) Error between subjective responses from Experiment 2 and a set of audio/speech quality metrics.

Additionally, Figure 4 (a) and (b) depict bar plots presenting the Root Mean Square Error (RMSE) between the subjective responses from Experiment 1 and 2 and the corresponding objective metrics. For the visual objective metrics (Figure 4 (a)), it can be observed that SSIM, DIVIINE, VIIDEO, NIQE, and NAVE presented lower error values. As for the audio objective metrics (Figure 4 (b)), P.563 and AQUA presented the lower error values.

Figure 5 (a) depicts the objective scores obtained with the video metrics versus the subjective scores from experiment 3. In this case, results are sparser, when compared to experiment

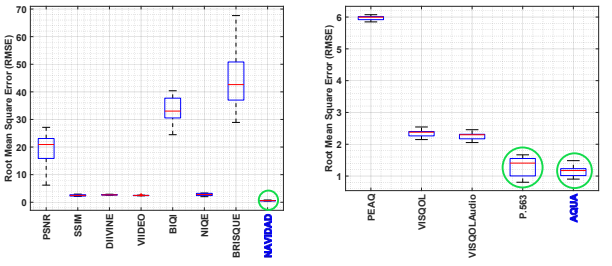
1. NAVE showed a much better correlation with the subjective scores than the rest of the metrics. Moreover, DIIVINE, VIIDEO, and NAVIDAD showed a fair correlation with the audio-visual subjective scores. Meanwhile, Figure 5 (b) shows the objective scores obtained with the audio quality metrics versus the subjective scores from experiment 3. Notice that VISQOL, VISQOLAudio and PEAQ had similar performances, when compared to results of experiment 2. As for P.563, AQUA and NAVIDAD, their predicted scores are more correlated with the subjective scores.




(a) (b)

Figure 5: Scatter Plot presenting a Subjective-Objective comparison. (a) Subjective responses from Experiment 3 against a set of visual quality metrics from literature. (b) Subjective responses from Experiment 3 against a set of audio/speech quality metrics.

Figure 6 (a) and (b) depict bar plots presenting RMSE between the subjective responses from Experiment 3 and the corresponding objective metrics for video and audio. For the visual objective metrics (Figure 6 (a)), it can be observed that SSIM, DIVIINE, VIIDEO, NIQE, and NAVIDAD presented lower error values. For the audio objective metrics (Figure 6 (b)), P.563 and AQUA presented the lower error values. These values are in agreement with the error results from experiments 1 and 2.




(a) (b)

Figure 6: Bar Plot presenting Root Mean Square Error (RMSE) comparison. (a) Error between subjective responses from Experiment 3 and a set of visual quality metrics from literature. (b) Error between subjective responses from Experiment 3 and a set of audio/speech quality metrics.

## Conclusions

This paper presents a study of the performance of several single and multi-modal quality metrics, when tested on a set of audio-visual content dataset, with impairments on both audio and video components. The goal was to test the performance of these metrics on a multimodal (audio and video) content. A number of audio and video objective quality metrics were gathered from the literature and they were tested against a set of metrics based on an autoencoder machine learning architecture. Overall, the autoencoder-based metrics presented better correlations with the subjective scores. They also presented low error rates for all the datasets. These results proved the value and capacity of the autoencoder-based metrics to predict the quality of signals. Further experiments are encouraged in order to improve the performance of the models.

## References


[1] Akhtar, Z. and Falk, T. H., "Audio-visual multimedia quality assessment: A comprehensive survey," *IEEE Access* **5**, 21090–21117 (2017).

[2] Garcia, M. and Raake, A., "Impairment-factor-based audio-visual quality model for iptv," in *Int. Workshop on Quality of Multimedia Experience (QoMEx), 2009.*, 1–6, IEEE (2009).

[3] Paolo Gastaldo and Judith A Redi. Machine learning solutions for objective visual quality assessment. In *6th international workshop on video processing and quality metrics for consumer electronics, VPQM*, volume 12, 2012.

[4] Helard A Becerra Martinez and Mylene CQ Farias. Using The Immersive Methodology to Assess The Quality of Videos Transmitted in UDP and TCP-Based Scenarios. *Electronic Imaging*, pages 233–1, 2018.

[5] Helard A Becerra Martinez and Mylene CQ Farias. Analyzing the influence of cross-modal IP-based degradations on the perceived audio-visual quality. *Electronic Imaging*, pages 324–1, 2019.

[6] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1398–1402. Ieee, 2003.

[7] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *Image Processing, IEEE Transactions on*, 25(1):289–300, 2016.

[8] Yi Zhang, Anush K Moorthy, Damon M Chandler, and Alan C Bovik. C-diivine: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes. *Signal Processing: Image Communication*, 29(7):725–747, 2014.

[9] AK Moorthy and AC Bovik. A modular framework for constructing blind universal quality indices. *IEEE Signal Processing Letters*, 17, 2009.

[10] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a" completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

[11] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[12] C. Sloan, N. Harte, D. Kelly, A. C. Kokaram, and A. Hines, "Objective assessment of perceptual audio quality using visqolaudio," *IEEE Transactions on Broadcasting*, vol. 63, no. 4, pp. 693–705, 2017.

[13] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes. Peaq-the itu standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29, 2000.

[14] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Visqol: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 13, 2015.

[15] L. Malfait, J. Berger, and M. Kastner, "The itu-t standard for single-ended speech quality assessment," Tech. Rep. 6, 2006.

[16] Helard A Becerra Martinez, Mylene CQ Farias, and Andrew Hines.

A No-Reference Autoencoder Video Quality Metric. *International Conference on Image Processing (ICIP) IEEE*, pages 1755–1759, 2019.

[17] Helard A Becerra Martinez. A three layer system for audio-visual quality assessment. *PhD Thesis*, 2019.

[18] Helard A Becerra Martinez, Mylene CQ Farias, and Andrew Hines. NAViDAd: A No-Reference Audio-Visual Quality Metric Based on a Deep Autoencoder. *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019.

## Author Biography

*Helard B. Martinez received his BS degree in computer science from Universidad Nacional San Antonio Abad del Cusco (UNSAAC), Peru, in 2010, his MSc degree in computer science from University of Braslia (UnB), Brazil, in 2013, and his PhD degree in computer science from University of Braslia (UnB), Brazil, in 2019. He was a visiting researcher at University College Dublin (UCD), Ireland, in 2017. His current research interests include signal processing, quality assessment metrics, and quality of experience.*

*Mylène C. Q. Farias received the B.Sc. degree in electrical engineering from the Universidade Federal de Pernambuco, Recife, Brazil, in 1995, the M.Sc. degree in electrical engineering from the Universidade Estadual de Campinas, São Paulo, Brazil, in 1998, and the Ph.D. degree in electrical and computer engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2004. She was previously a Research Engineer with CPqD, Campinas, Brazil, as well as an Intern with Philips Research Laboratories, Eindhoven, The Netherlands, and with the Intel Corporation, Phoenix, AZ, USA. She is currently an Associate Professor with the Department of Electrical Engineering, University of Brasília, Brasília, Brazil. Her current interests include video quality metrics, video processing, multimedia, watermarking, and machine learning.*

*Andrew Hines is an Assistant Professor with the School of Computer Science, University College Dublin, Ireland. He leads the QxLab research group and is an investigator at the SFI CONNECT Centre for Future Networks and SFI Insight Centre for Data Analytics. His primary research interests are in media signal processing and machine learning for data driven quality of experience prediction across a variety of domains.*

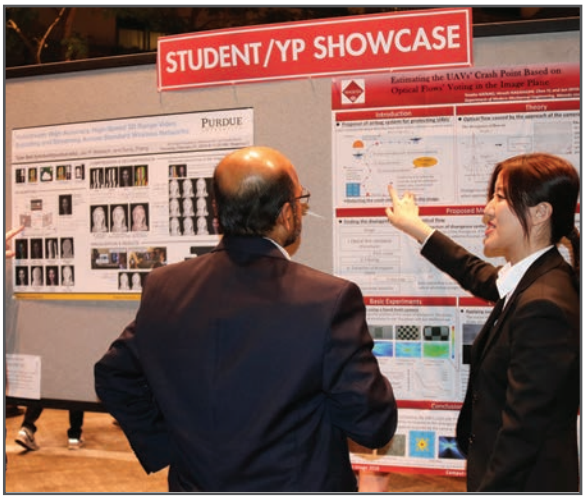
# JOIN US AT THE NEXT EI!

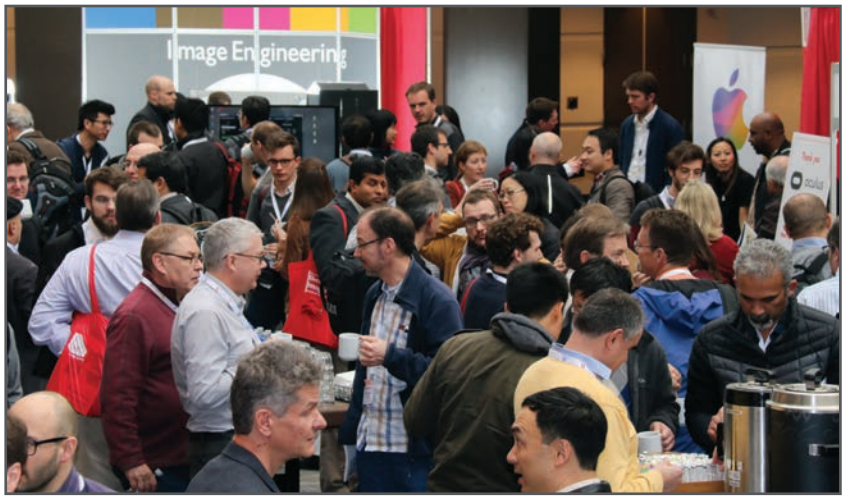
IS&T International Symposium on

# 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