https://doi.org/10.2352/ISSN.2470-1173.2020.11.HVEI-068

Improved Temporal Pooling for Perceptual Video Quality Assessment Using VMAF

Sophia Batsi and Lisimachos P. Kondi;
Department of Computer Science and Engineering; University of Ioannina, Ioannina, Greece

Abstract

The Video Multimethod Assessment Fusion (VMAF) method, proposed by Netflix, offers an automated estimation of perceptual video quality for each frame of a video sequence. Then, the arithmetic mean of the per-frame quality measurements is taken by default, in order to obtain an estimate of the overall Quality of Experience (QoE) of the video sequence. In this paper, we validate the hypothesis that the arithmetic mean conceals the bad quality frames, leading to an overestimation of the provided quality. We also show that the Minkowski mean (appropriately parametrized) approximates well the subjectively measured QoE, providing superior Spearman Rank Correlation Coefficient (SRCC), Pearson Correlation Coefficient (PCC), and Root-Mean-Square-Error (RMSE) scores.

Introduction

Quality of Experience (QoE) is a performance metric that focuses on the customer and refers to his/her experience and satisfaction from a service. QoE can be measured by gathering human ratings in a subjective quality evaluation test. In this case, the Mean Opinion Score (MOS) is a widely used measure. MOS scores can also be predicted by objective quality metrics which typically have been developed and trained using human ratings. Video Multimethod Assessment Fusion (VMAF) is a video quality metric developed by Netflix to predict the subjective video quality based on MOS. Estimations of quality on a per-frame basis are used for the production (through a temporal pooling method) of a summary score for the whole video.

Results in the literature have already proven that the mean as a pooling method conceals the frames with large degradation, giving equal weight to all the frames' quality scores without consideration of the distortion level of the frames. For example, in [1] the authors compare (in terms of PSNR, SSIM, VQM, MSE and sqrtMSE) the performance of a set of pooling methods and claim that pooling methods which give more weight to the most recent (recency effect) or most distorted frames of a video, perform best. In [2], several temporal pooling methods are compared for per-frame quality measures such as PSNR and SSIM. Authors in [3] present the correlation of PSNR with MOS, after using different pooling methods.

However, none of the above works consider VMAF. Thus, there is still no clear direction in the literature regarding the best choice for the temporal pooling method to be used when VMAF is applied. To fill this gap, we applied the Minkowski pooling method, $k-th$ percentile, and mean last frames with VMAF on Netflix recommended datasets [4, 13] and conducted a set of tests measuring Spearman Rank Correlation Coefficient (SRCC), Pearson Correlation Coefficient (PCC) and Root-Mean-Square-Error (RMSE).

The VMAF Metric

Video streaming providers try to offer the best quality experience to their customers. For this reason, a new video quality metric has been proposed, called Video Multimethod Assessment Fusion (VMAF) [4]. It is a full reference video quality metric that aims to approximate human perception provided in terms of MOS or Differential Mean Opinion Score (DMOS). VMAF extracts on a per-frame basis the following elementary features:

- the Detail Loss Measure (DLM) [5],
- the Visual Information Fidelity, VIF [6] and
- the luminance difference between pairs of frames (Temporal Information)

For the training of VMAF, the values of the above features are calculated for each frame of the training video sequences. Then, the arithmetic mean of each feature is taken over the whole video sequence. The video quality ground truth is the MOS or DMOS for the whole sequence, obtained from experiments with human observers. Clearly, it is very hard, if not impossible, to obtain per-frame ground truth from experiments with humans. Thus, the average of the features together with the MOS or DMOS are fed to a Support Vector Machine (SVM) [7, 8] model.

In testing, the feature values for each frame of the video sequence are input to the SVM model to output the estimated video quality of the frame. By default, VMAF uses the arithmetic mean of all the frames' scores to provide an overall video quality score [1]. As it has been proven for specific video data sets, VMAF scores have stronger correlation to subjective MOS, compared to other quality metrics such as PSNR, and SSIM [9].

Temporal Pooling Methods

The procedure for computing the VMAF score of a video, as mentioned in [10], consists of the following steps: feature extraction and aggregation, training/testing, and temporal pooling. Temporal pooling refers to the method in which a series of frame quality scores result in one quality score for the whole video sequence. This can be achieved by a variety of temporal pooling methods.

In this paper, three temporal pooling methods are used: Minkowski summation, $k-th$ percentile and mean value of scores in last F frames.

The Minkowski summation is given by the formula

$$\boldsymbol{OM_{Mink}} = \left[\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{OM^{P}(t)}\right]^{1/p} \quad \textbf{(1)}$$

where T is the number of frames in a video (frame sequence), p is the Minkowski exponent and OM stands for VMAF scores. As p values increase, the influence of high-quality frames is emphasized. As we can see in Eq. (1), for different values of p we have different temporal pooling methods.

The mean value of frames' scores is given if we set the exponent to $p = 1$ in Eq. (1):

$$OM_{mean} = \sum_{t=1}^{T} OM(t), \quad (2)$$

where T and OM have the same meaning as in Eq. (1).

For $p = -1$ in (1) we have the harmonic mean [11]. The harmonic mean often produces a summary score very similar to the mean, except that in the presence of outliers, the harmonic mean emphasizes the impact of small values:

$$OM_{Harmonic} = \left[\frac{1}{T}\sum_{t=1}^{T} OM^{-1}(t)\right]^{-1}. \quad (3)$$

With another choice of the Minkowski's summation exponent, $p = 2$, Eq. (1) becomes the quadratic mean, the square root of the arithmetic mean of the squares of the per frame values, also known as Root Mean Square :

$$OM_{RMS} = \left[\frac{1}{T}\sum_{t=1}^{T} OM^{2}(t)\right]^{1/2}. \quad (4)$$

The Mean Last Frames pooling method computes the mean quality score of most recent F frames.

$$OM_{meanF} = \frac{1}{F}\sum_{t=T-F}^{T} OM(t), \quad (5)$$

High values of F result in weaker recency effect.

$K - th$ percentile [12] of an ordered set is the lowest k% values of a set. Low values of k show the influence on users of the lowest quality frames.

Experimental Results

To evaluate the performance of different pooling methods, we test them on the two main datasets for which the VMAF metric has already been validated, i.e., the NETFLIX Video dataset and the Video Quality Expert Group HD3 (VQEG HD3) dataset [13].

From the NETFLIX Video Dataset, we used nine six-second reference videos with both high level and low-level characteristics. For each original video, distorted videos have been produced, encoded H.264/AVC video streams at resolutions between 384×288 to 1920×1080 and bitrates between 375 kbps to 20000 kbps. There are a total of 70 distorted videos. Each of the videos, have been exposed to subjective test, acquiring a Differential Mean Opinion Score (DMOS score) normalized between 1 and 100.

From the VQEG HD3 dataset we used a subset of eight reference videos and for each one of them, eight distorted videos were produced with two types of encoding: MPEG-2 and H.264, 64 distorted in total, 10 seconds long each.

As described above, most of the temporal pooling methods require input parameters. For this reason, all parametric temporal pooling algorithms are tested for several parameters values in order to find the optimal value of a temporal pooling parameter, which maximize the correlation between the temporal pooling method and the subjective scores.

We used three optimization criteria: The Pearson Correlation Coefficient (PCC), the Spearman Rank Correlation Coefficient (SRCC) and the Root-Mean-Square-Error (RMSE). The SRCC measures the monotonic relationship between the objective predictions and the subjective scores, while PCC measures the degree of linearity between the two. Both correlation coefficients describe the overall agreement between objective scores and ground truth. Values closer to 1 are more desirable. RMSE values closer to 0 will indicate a perfect fit between results and objective scores.

In Figures 1 to 3 the SRCC, PCC and RMSE for VMAF using Minkowski summation as pooling method are displayed respectively. Values -1, 0.5, 1, 2, 2.5, 3, 3.5, 4, 5, 8, 10, 50 and 100 were selected for Minkowski exponents. The best results were achieved for $p = 8$. As p increased from -1 to 8, the correlation between the VMAF score and MOS was stronger. After $p = 8$ the correlation between the two scores starts to decline.

For the Mean Last Frames pooling method, Figures 4 to 6 show respectively that by taking the arithmetic mean of the last 50 frames gives the best results for SRCC, PCC and RMSE for the NETFLIX Video Dataset, however for the VGEG HD3 the best results were achieved when $F = 100$. For the F parameter in Eq. (5) we considered the values 25, 50, 75 and 100.

Figures 7, 8 and 9 display the SRCC, PCC and RMSE for the $k - th$ percentile. As k values, we studied 5, 10, 20 and 25. From the above, we come to the conclusion that for $k = 25$ we have the best results for both datasets.

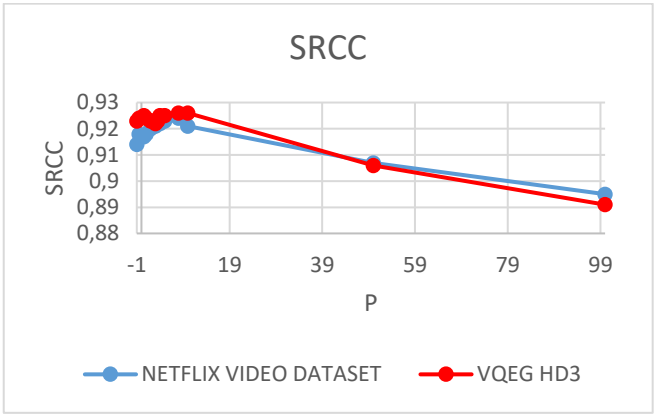


Fig. 1: Spearman Correlation Coefficient for VMAF using Minkowski summation as pooling method with different Minkowski exponents p

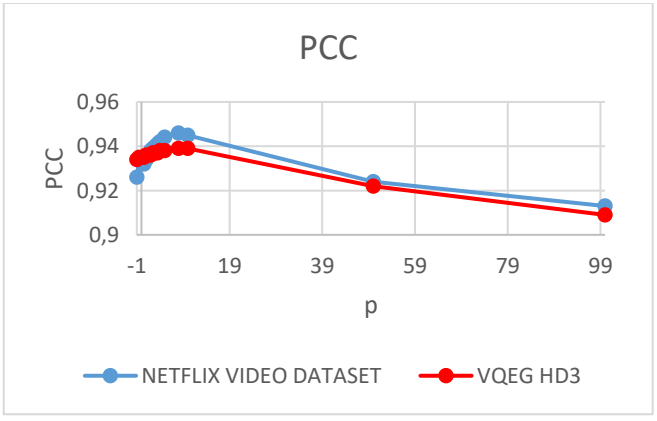


Fig. 2: Pearson Correlation Coefficient for VMAF using Minkowski summation as pooling method with different Minkowski exponents p.

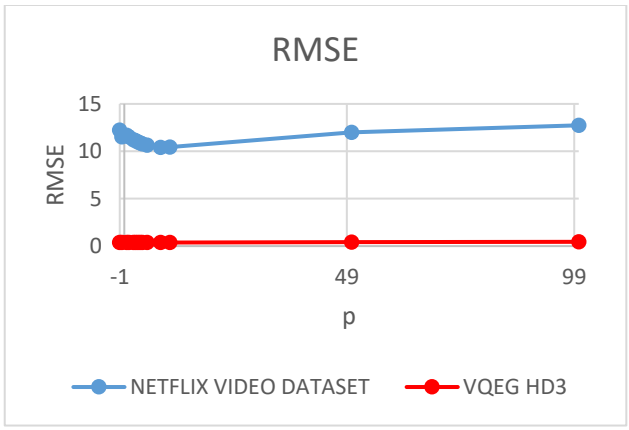


Fig. 3: Root Mean Square Error for VMAF using Minkowski summation as pooling method with different Minkowski exponents p.

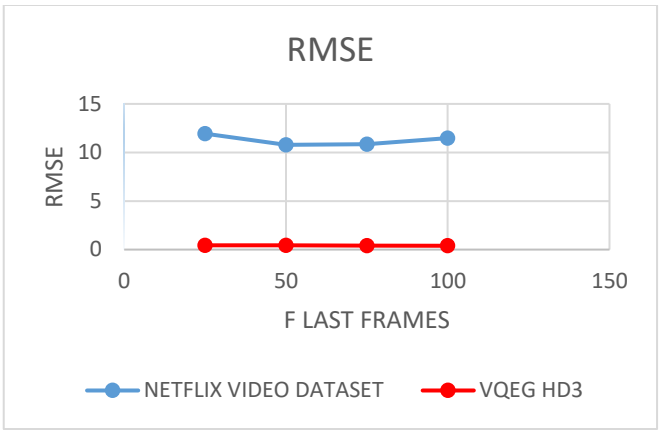


Fig. 6: Root Mean Square Error Coefficient for VMAF using Mean Last Frames as pooling method with different F.

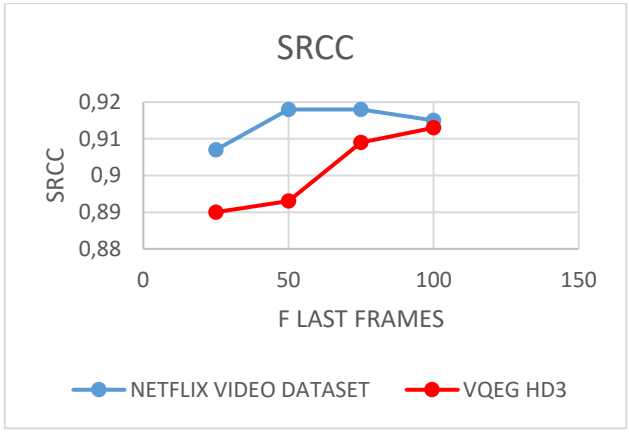


Fig. 4: Spearman Correlation Coefficient for VMAF using Mean Last Frames as pooling method with different F.

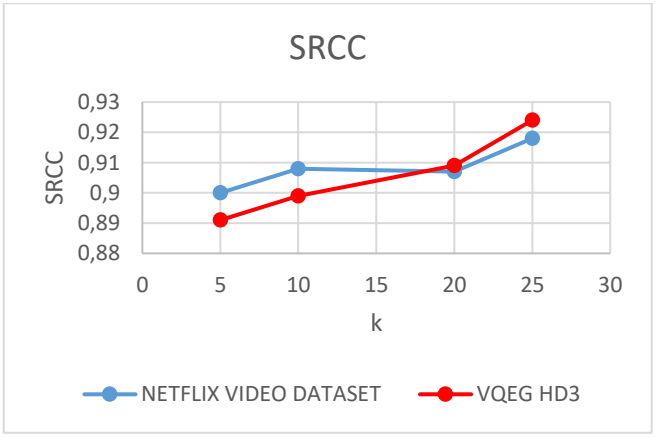


Fig. 7: Spearman Correlation Coefficient for VMAF using $k-th$ percentile as pooling method with different k.

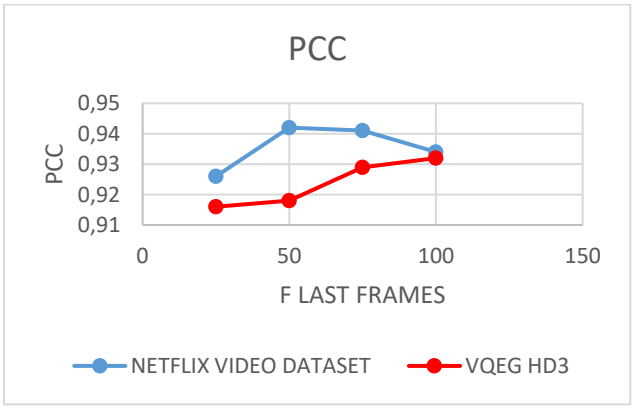


Fig. 5: Pearson Correlation Coefficient for VMAF using Mean Last Frames as pooling method with different F.

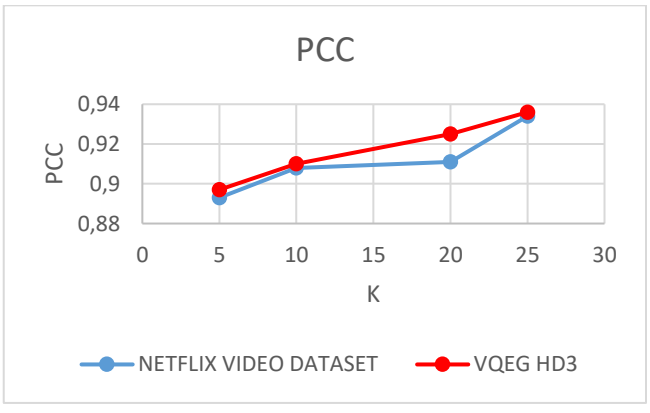


Fig. 8: Pearson Correlation Coefficient for VMAF using $k-th$ percentile as pooling method with different k.

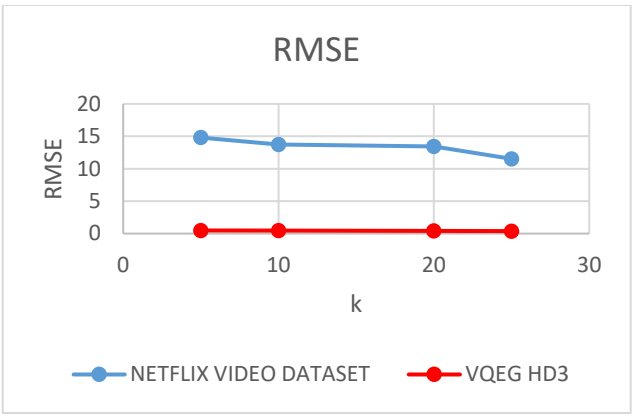


Fig. 9: Root Mean Square Error Coefficient for VMAF using $k-th$ percentile as pooling method with different k.

For the sake of comparison, in Table 1, the performance of all the pooling methods used, is provided. As can be observed, for the NETFLIX Video Dataset and the VQEG HD3 Dataset the best performing method was the Minkowski summation with $p = 8$.

Still, Minkowski with $p = 5$, Minkowski with $p = 10$, Mean of Last Frames with F= 50 and Quadratic Mean, performed better than the arithmetic Mean in NETFLIX Video Dataset, and Minkowski with $p = 5$ and $p = 10$ performed better than arithmetic Mean in VQEG HD3.

This shows that pooling methods which give more weight on frames with high distortion correlate better with subjective scores.

Table 1. Performance comparison of temporal pooling methods in NETFLIX Video dataset and in VQEG HD Dataset.

		VMAF					
Pooling Method		NETFLIX DATASET			VQEG HD3		
		SRCC	PCC	RMSE	SRCC	PCC	RMSE
Minkowski	$p = -1$	0.914	0.926	12.249	0.923	0.934	0.387
	$p = 1$ (VMAF default)	0.918	0.934	11.529	0.924	0.936	0.384
	$p = 2$	0.920	0.938	11.234	0.923	0.936	0.383
	$p = 5$	0.923	0.944	10.642	0.925	0.938	0.38
	$p = 8$	**0.924**	**0.946**	**10.422**	**0.926**	**0.939**	**0.377**
	$p = 10$	0.921	0.945	10.434	0.926	0.939	0.376
Last Frames	$F = 20$	0.907	0.926	11.937	0.890	0.916	0.439
	$F = 50$	0.918	0.942	10.785	0.893	0.918	0.432
	$F = 100$	0.915	0.934	11.482	0.913	0.932	0.392
Percentile	$k = 5$	0.900	0.893	14.803	0.891	0.897	0.482
	$k = 10$	0.908	0.908	13.741	0.899	0.91	0.452
	$k = 20$	0.907	0.911	13.422	0.909	0.925	0.41
	$k = 25$	0.918	0.934	11.528	0.924	0.36	0.384

Conclusions

By comparing three different temporal pooling methods for calculating the VMAF score of a video sequence on the NETFLIX Video Dataset and the VQEG HD3 Dataset, we conclude to the importance of the choice of a pooling method. Different pooling methods can remarkably change the VMAF score. The best results have shown to us the well-known fact that QoE scores resulting from users scoring, are influenced the most by two things: The most degraded part of the video and the quality of it in the last seconds of the video.

Hence, for results that correspond to human perception quality of a video we should resort to pooling methods which have those two characteristics. Our experiments showed that the Minkowski pooling method (appropriately parametrized) outperform all other pooling methods.

From the study presented in this paper, it is revealed that a deeper investigation on a potential relation between the video types/content and the pooling method selected, is worth to be conducted.

References

[1] S. Rimac-Drlje, M. Vranjes, and D. Zagar, “Influence of temporal pooling method on the objective video quality evaluation,” IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, 2009.

[2] M. Seufert, M. Slanina, S. Egger, M. Kottkamp, “To pool or not to pool: A comparison of temporal pooling methods for HTTP adaptive video streaming”, *Fifth International Workshop on Quality of Multimedia Experience (QoMEX),* Klagenfurt, Austria, July 2013.

[3] C. Keimel and K. Diepold, “Improving the prediction accuracy of PSNR by simple temporal pooling,” *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Jan. 2010.

[4] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, “Toward a practical perceptual video quality metric [Online]. Available: http://techblog.netflix.com/2016/06/towardpractical-perceptual-video.html

[5] S. Li, F. Zhang, L. Ma, and K. N. Ngan, “Image quality assessment by separately evaluating detail losses and additive impairments,” *IEEE Trans. Multimedia,* vol. 13, no. 5, pp. 935–949, 2011.

[6] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Trans. Image Process.,* vol. 15, no. 2, pp. 430–444, 2006.

[7] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning,* vol. 20, no. 3, pp. 273–297, 1995.

[8] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intelligent Systems and Technology,* vol. 2, no. 3, May 2011.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Trans. Image Process.,* vol. 13, no. 4, pp. 600–612, 2004.

[10] C. G. Bampis, Z. Li, and A. C. Bovik, "SpatioTemporal feature integration and model fusion for full reference video quality assessment," in arXiv:1804.04813 e-print, 2018.

[11] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy and J. De Cock, “VMAF: The Journey Continues”. [Online]. Available: https://medium.com/netflix-techblog/vmaf-the-journey-continues-44b51ee9ed12

[12] A. K. Moorthy and A. C. Bovik, “Perceptually significant spatial pooling strategies for image quality assessment,” *SPIE Human Vis. Electron. Imag.,* vol. 7240, pp. 724012-1-–724012-11, Jan. 2009.

[13] HDTV Phase I Final Report [Online] Available: https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx

Author Biography

Sophia Batsi holds a BSc and an MSc in Computer Science from the University of Ioannina (2016 and 2019, respectively). Her research work is focused on perceptual video quality.

Lisimachos P. Kondi received the PhD degree in electrical and computer engineering from Northwestern University, Evanston, IL, USA, in 1999. He is currently Professor in the Department of Computer Science and Engineering, University of Ioannina, Greece. His research interests are in the general areas of signal and image processing and communications, including image and video compression and transmission over wireless channels and the Internet, sparse representations and compressive sensing, and super-resolution of video sequences.

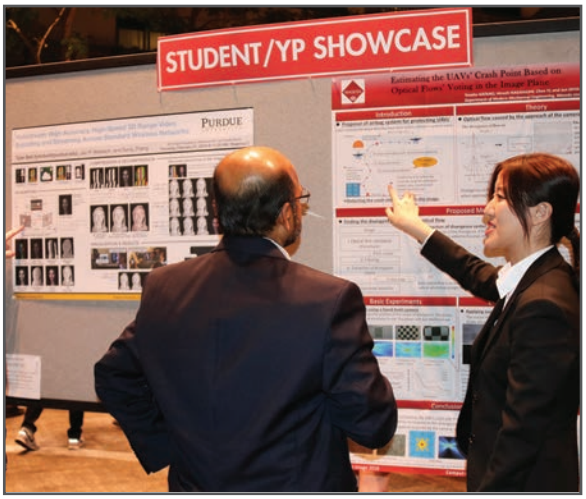
JOIN US AT THE NEXT EI!

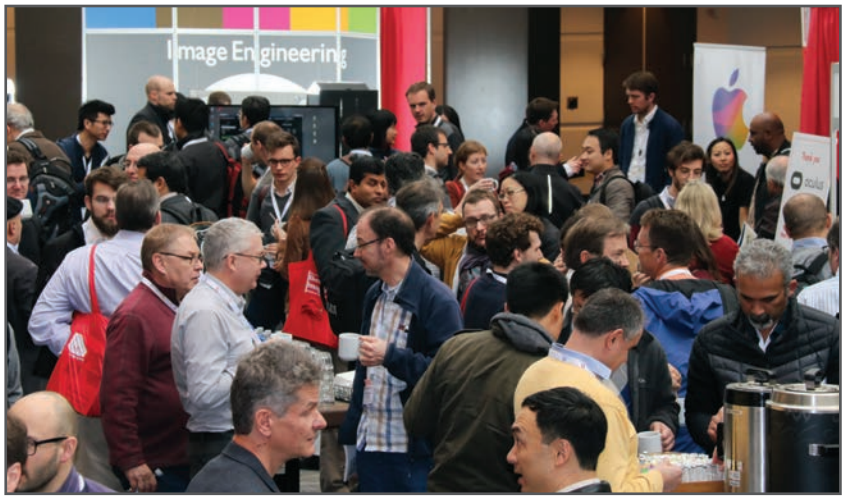
IS&T International Symposium on

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