Investigating Prediction Accuracy of Full Reference Objective Video Quality Measures through the ITS4S Dataset

https://doi.org/10.2352/ISSN.2470-1173.2020.11.HVEI-093

Investigating Prediction Accuracy of Full Reference Objective Video Quality Measures through the ITS4S Dataset

Lohic Fotio Tiotsop, Antonio Servetti, Enrico Masala
Control and Computer Engineering Department
Politecnico di Torino, Torino, Italy

Abstract

Large subjectively annotated datasets are crucial to the development and testing of objective video quality measures (VQMs). In this work we focus on the recently released ITS4S dataset. Relying on statistical tools, we show that the content of the dataset is rather heterogeneous from the point of view of quality assessment. Such diversity naturally makes the dataset a worthy asset to validate the accuracy of video quality metrics (VQMs). In particular we study the ability of VQMs to model the reduction or the increase of the visibility of distortion due to the spatial activity in the content. The study reveals that VQMs are likely to overestimate the perceived quality of processed video sequences whose source is characterized by few spatial details. We then propose an approach aiming at modeling the impact of spatial activity on distortion visibility when objectively assessing the visual quality of a content. The effectiveness of the proposal is validated on the ITS4S dataset as well as on the Netflix public dataset.

Introduction

It is well known that to perform reliable studies in the field of subjective video quality estimation, large subjectively annotated datasets are needed. Recent years witnessed an increasing amount of publicly available datasets of source sequences for video quality research purposes. However, finding large datasets, in terms of number of unique source sequences, with good content variety and with reliable subjective quality measures is still difficult. In this work we focus on a recently released dataset from NTIA/ITS [1], named ITS4S, that contains more than 800 unique source sequences, with the corresponding subjective annotation. In this work we augmented such datasets by adding eight different objective video quality measures (VQM) for each one of the subjective evaluations present in such dataset. Moreover, in addition to the video sequence classification that has already been done by the dataset designers into 9 topic categories, we also computed two objective indexes that are typically used to characterize video sequences, namely the spatial and temporal complexity indexes, SI and TI. After computing the value of these features for each frame, we obtained the final value for a given source by using three different pooling operators; taking the maximum value (maxSIsrc, maxTIsrc), the average (avgSIsrc, avgTIsrc) and finally the minimum value (minSIsrc, minTIsrc).

The newly added part of the dataset is available at http://media.polito.it/its4s free to use by any researcher working in the objective video quality measures field. In this work we first study the dataset characteristics. In particular, relying on statistical methods, we study the diversity of the sequences in the dataset in terms of the ability of full reference VQMs to accurately predict their perceived visual quality. The goal of such analysis is to determine whether the content of the dataset is enough heterogeneous and thus how suitable it could be for training and/or investigating the accuracy of VQMs. This is particularly useful when case machine learning (ML) based approaches are used. For instance, the success of ML based measures is strongly related to the amount of information that could be learned from the training set. Hence, determining how informative, i.e., how diverse, is the content in a dataset according to some criteria related to quality perception is of paramount importance.

Relying on the ITS4S extension we investigate the reliability of widely used full reference objective measures when predicting subjective quality as a function of the spatial activity of the content. The aim is to determine whether full reference VQMs can correctly model the emphasis of distortion visibility due to low spatial activity. In fact the visibility of the distortion could be significantly affected by the quantity of spatial details in the content [2]. Distortion tends to be less visible in presence of high spatial activity or motion while the absence of details could significantly emphasize the distortion visibility.

The contribution of this work can be summarized as follows: i) an extension of the ITS4S dataset is made available by adding objective measures as well as features extracted from the sources and processed video sequences; ii) investigating the characteristics of the dataset content in terms of heterogeneity and suitability for the design, validation and testing of objective metrics; iii) studying and modeling the impact of the spatial activity on the accuracy of full reference VQMs when predicting the quality as perceived by human observers;

The ITS4S Dataset: Description and Extension

The ITS4S dataset was build and published by NTIA/ITS [1] in the CDVL repository [3]. The initial aim of the dataset was to run a subjective experiment to gain preliminary insight on the possibility to develop reliable no-reference objective measures. However the dataset as well as the related subjective scores can also be used for other research purposes, as done in the following contribution. For further information regarding the dataset as well as the original use case, the reader may refer to the following report [4].

There are 813 unique source sequences in the ITS4S dataset, each about 4 second long. Despite the rather limited time duration, they have been shown to be useful in subjective quality evaluation experiments [4]. Each of these sequences was either left unaltered or processed using one out of five hypothetical reference

circuits (HRCs), that is, compressing the video using the AVC standard High Profile at one of these 5 different bitrate values: 512, 951, 1256, 1732, 2340 kbps. The original content resolution is 1280x720. Such resolution has been used for the 2340 kbps encoding, whereas for lower bitrates, downsampling has been used before encoding, down to 1024x576, 824x464, 696x392, 512x288, respectively. However, note that all content has been decoded and upscaled again at 1280x720 before performing any subjective evaluation. The subjective experiment used a modified version of the ACR rating scale, where in addition to the standard five level scale (excellent, good, fair, poor, bad), two other responses were available, in which the subject stated that it was not able to perform the evaluation due to some issues (e.g., distraction or computer glitch).

In addition to releasing all the processed video sequences (PVS) used in the experiment, the authors also publicly released all the original uncompressed source sequences. Therefore, this allowed us to run full reference objective measures in order to obtain the objective quality on many of the PVS. In fact, the original experiment asked the subjects to rate the quality of both unprocessed sequences (e.g., not subject to any of the previous HRCs) and PVSs resulting from the application of a given HRC to a source sequence. Therefore, in the latter case, both the source sequence (SRC) and the PVS was available, so we have been able to compute objective VQMs on such pairs. In the end, there are 514 available pairs that we used in our extended dataset. For each one of them, after an initial temporal alignment step that required a constant shift for all sequences equal to 1 frame, we computed the following 8 objective measures: PSNR, SSIM, MS-SSIM, VIF as implemented by the VQMT software [7], two versions of the VMAF measure [8] and two other variants of the SSIM and MS-SSIM measures as implemented by Netflix vmaf software [8]. They are named $SSIM_v$ and $MS\text{-}SSIM_v$ in the reminder of the work. For the case of VMAF, we computed the value resulting from version 0.6.2 of the VMAF model. In addition, we also considered the VMAF value estimated by the same version of the model through the bootstrap aggregation technique, as implemented by the VMAF software [8]. This is referred to as $VMAF_b$ in the remainder of this work. Finally, in addition to the previous VQM measures, for each SRC we also computed an index of spatial perceptual information (SI) and temporal perceptual information (TI), borrowing the indexes as defined by the ITU-T Rec. P.910 [5]. The indexes have been computed using the SITI software [9]. Such measures allowed us to both characterize the SRC sequences in an objective manner, as well as to investigate correlations between spatial and temporal activity and the accuracy of the predicted subjective scores by the objective measures.

The sequences in the ITS4S dataset have been manually divided by the dataset creators into nine categories, namely Broadcast, Chance, Everglades, Music&Mexico, Nature, Ocean, Public Safety, Sports, Training. A more detailed description of the content in available in [4]. Here we only note that the Chance category contains miscellaneous content that would not fit in the other categories.

Content Diversity and Patterns Investigation

This section is devoted to show that the sequences in the ITS4S dataset are characterized by a strong heterogeneity by the point of view of quality assessment and thus such dataset can be considered an effective training and test set for the development and/or the validation of new VQMs.

Contingency table.

	vqm			
category	VMAF	PSNR	SSIM	VIF
Broadcast	40	34	32	46
Chance	34	27	30	26
Everglades	45	23	31	36
Music&Mexico	40	28	23	30
Nature	48	28	27	36
Ocean	32	23	26	38
Public Safety	43	31	36	50
Sports	43	29	29	45
Training	5	3	3	5
Total	330	226	237	312
Frequency	64.2%	43.9%	46.1%	60.7%

We start by performing a least square fit of the scores of each VQM to the MOS scale using the equation

$$\widehat{VQM} = \widehat{P}_{VQM}(VQM) \tag{1}$$

where $\widehat{P}_{VQM}$ is a polynomial function as recommended in [6] and $\widehat{VQM}$ represents the MOS estimation.

Our approach relies on two main statistical tools: the confidence interval (CI) and the correspondence analysis (CA). While the CI is well know and widely used, we are not aware of other works that use the correspondence analysis to visualize and thus recognize interesting patterns in video quality assessment (VQA). The main purpose of the CA is to represent, in a 2D plot, the degree of association or dissociation of the values of two different categorical variables. The method allows to reproduce in 2D, thus easily interpret, some information contained in the contingency table of two categorical variables. The reader is referred to [10] for more details about the CA.

To perform the CA two categorical variables are required: in our case we use a variable *vqm* which can assume one of the values VMAF, PSNR, SSIM, VIF and another variable *category* that

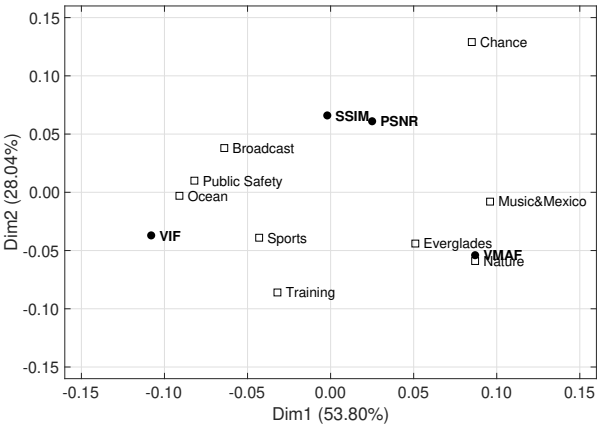


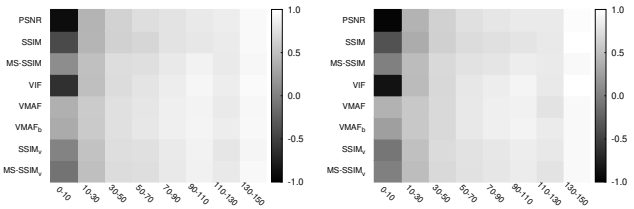
Figure 1. *The output of the correspondence analysis performed on the ITS4S dataset.*

can assume values equal to any one of the categories included in the dataset. We experimented also with more values for the *vqm* variable, but this only increased the difficulty of interpreting the results of the analysis without providing any additional insight. We build a contingency table for the two variables as follows. For each PVS we computed the 95% CI of its MOS, then we counted for each VQM the number of PVSs such that the VQM provided a fitted $\widehat{VQM}$ score in the 95% CI of the MOS. Finally for each PVS such that none of the VQM provided a prediction in the corresponding CI, we looked for the VQM with nearest score to its CI and incremented the number of PVSs associated to that VQM by 1. In other words, for each measure we count the number of PVSs of each category whose predicted objective score cannot be said to be different from the MOS with 95% of confidence. Table 1 summarizes the results. The last line (frequency) reports the ratio between the sum of the occurrences and the total number of PVSs in the dataset (514). It can be observed that the value for VMAF is the highest with respect to the other VQMs. However, for up to 35.8% of the PVSs (the ones not counted in any of the categories in the table), the VMAF measure provided a score significantly different than the MOS with 95% of confidence.

In Table 1, each category of the dataset can be interpreted as a point in a 4-dimensional space generated by the VQMs and each VQM a point in the 9-dimensional space generated by the categories of the dataset. Unfortunately it is not possible to directly visualize data in 4D or 9D in order to identify interesting similarities or patterns. The CA allows us to find an optimal common 2D space in which it is possible to represent both the VQMs and the categories maintaining as much as possible the initial dispersion of the categories in the 4D space and that of the VQMs in the 9D space.

Table 1 can be given in input to any statistical software that automatically performs the CA shown in Figure 1. In our case we relied on "R" [11]. The percentages shown on the two axis (53.04% and 28.04%) indicate that moving from the original 9D spaces of the VQMs and 4D space of the categories to the common 2D representation, up to 81.84% (53.80%+28.04%) of the initial dispersion of the data is kept. The proximity between two values of the same variable on the graph determines how similar they are, i.e., it reveals the existence of patterns between two values associated to different variables. The analysis shows that, in general, content categories are rather widespread, suggesting that the dataset content is quite heterogeneous. However, some content is more similar, for instance the PVSs in the Nature and Everglades categories are very similar in terms of quality prediction and that the VMAF prediction on these categories is more precise than that of any other VQM. The same similarity is observed for the Ocean and Public Safety categories for which the analysis suggests using VIF as the best VQM. The Chance category does not seem to prefer any specific measure. This could be explained by the fact that the content of that category does not refer to a specific theme. We finally observe that the prediction capability of the PSNR and the SSIM on the dataset are rather similar.

Please note that the analysis conducted here allows to effectively sub-sample the data in the dataset should it be used to train an ML-based VQM. In fact any training made on a sub-sample composed of a majority of PVSs belonging to categories very similar to each other such as Nature and Everglades or Ocean and Public Safety would lead to VQM performing well on that sub-sample but probably not so well on other datasets.



(a) PLCC (b) SROCC

Figure 2. *The Pearson and Spearman rank correlation coefficient between the VQM scores and the MOS as a function of SI bands (shown on the horizontal axis.) The v subscript indicates the measures as computed by the VMAF software. The b subscript indicates the bootstrap aggregation version of VMAF.*

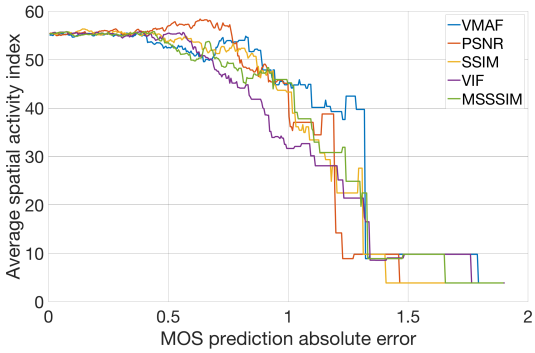


Figure 3. *Average spatial activity index as a function of the absolute value of the MOS prediction error. All VQMs show, on average, higher prediction error when predicting the quality of PVSs with low SI.*

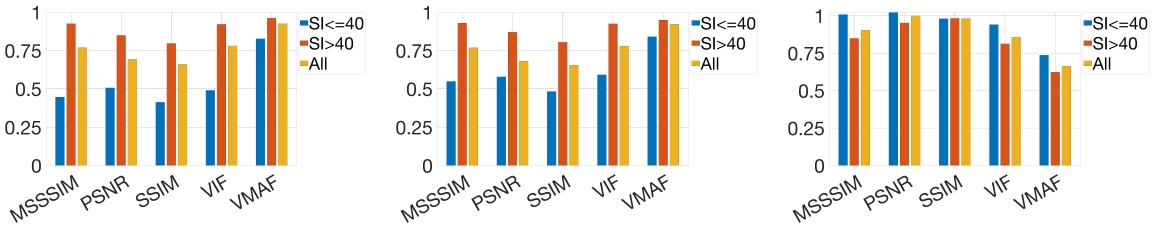
Spatial Activity and Perceived Visual Quality Prediction

In this section we study the influence of the spatial activity on the accuracy of the VQMs when predicting the visual quality of a content. In particular the analysis shows that the VQMs are likely to inaccurately predict the MOS of the PVSs derived from sources characterized low spatial activity. We first studied the correlation of the VQMs score to the MOS as function of the avgSIsrc, which seemed the most promising indexes among the ones added to the dataset. Figure 2 shows that VQMs are likely to deliver score poorly or even negatively correlated to the MOS in correspondence of sequences whose source is characterized by low values of avgSIsrc.

We then investigated also the behavior of the absolute prediction error of the MOS i.e., the difference between the MOS and the VQM prediction in absolute value, as function of the avgSIsrc. More precisely, for each VQM, we selected all the PVSs whose MOS prediction occurred with more than a given absolute error (x axis) then computed and represented the average of the avgSIsrc of those sequences (y axis).

Figure 3 shows that, on average, higher prediction errors occur in correspondence with sequences characterized by low spatial activity.

To ensure that the relation between the accuracy of the



(a) PLCC (b) SROCC (c) RMSE

Figure 4. *Correlation measures and RMSE between VQMs and MOS. VQMs show lower accuracy when assessing the quality of PVSs characterized by low SI, on the Netflix Public Dataset.*

VQMs and the spatial activity discussed so far is not just a peculiarity of the ITS4S dataset, we conducted some experiments also on the Netflix public dataset. We computed the Spearman rank order correlation coefficient (SROCC) between the VQMs scores and the MOS. In addition, after performing a polynomial fitting of all the VQMs to the MOS, we also computed the Pearson linear correlation coefficient (PLCC) and the residual mean square error (RMSE) separately for the sequences with $avgSIsrc \leq 40$ and those with $avgSIsrc > 40$. The results in Figure 4 further support the existence of a strong relation between the accuracy of the VQMs and the SI since higher correlation coefficient and lower RMSE values are observed for sequences with higher avgSIsrc value.

Modeling the Spatial Activity Influence on Distortion Visibility

In light of the results of the previous section, we believe it is worth modeling explicitly the influence of the avgSIsrc (referred to as SI in the following for simplicity's sake) when objectively predicting the perceived visual quality. In order to evaluate the accuracy of a specific VQM when used to assess the quality of a given PVS the following equation is usually employed:

$$MOS = f_{VQM}(VQM) + \varepsilon \tag{2}$$

where f_{VQM} is typically a polynomial or a logistic function (as recommended by [5]) and ε is a normally distributed error term.

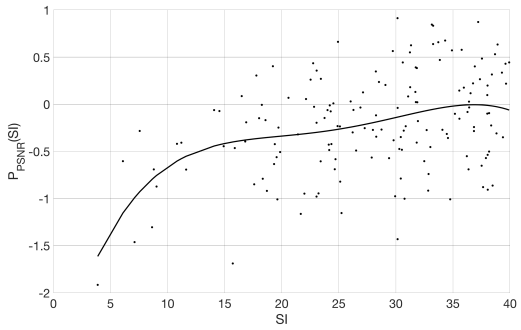
Since from the results of the previous section we observed that low values of SI seems to be related to a lower VQMs accuracy, in order to address such issue, we propose to improve the MOS predictions generated by the VQMs values by modeling the impact of the SI as follows:

$$MOS = f_{VQM}(VQM) + P_{VQM}(SI) \cdot \mathbf{1}_{SI<30} + \eta \tag{3}$$

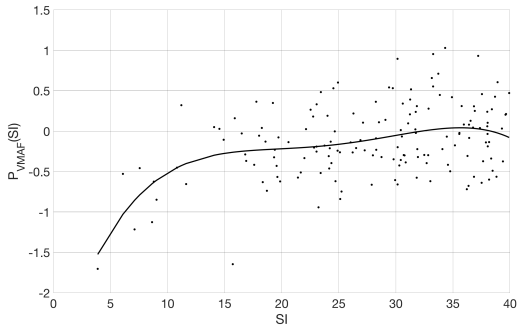
where $\mathbf{1}_{SI<30}$ is 1 when SI<30 and 0 otherwise, η is the residual error still present after the correction, and the correction term $P_{VQM}(SI)$ is a third order polynomial function whose coefficients are computed by performing a least square fitting of the residuals ε shown in the model in Eq. 2, as a function of the SI value.

Numerical Experiments

In order to demonstrate the effectiveness of the proposed approach i.e., modeling the influence of the SI through a third order polynomial function, we conducted numerical experiments aiming at providing further insight on the relation between the



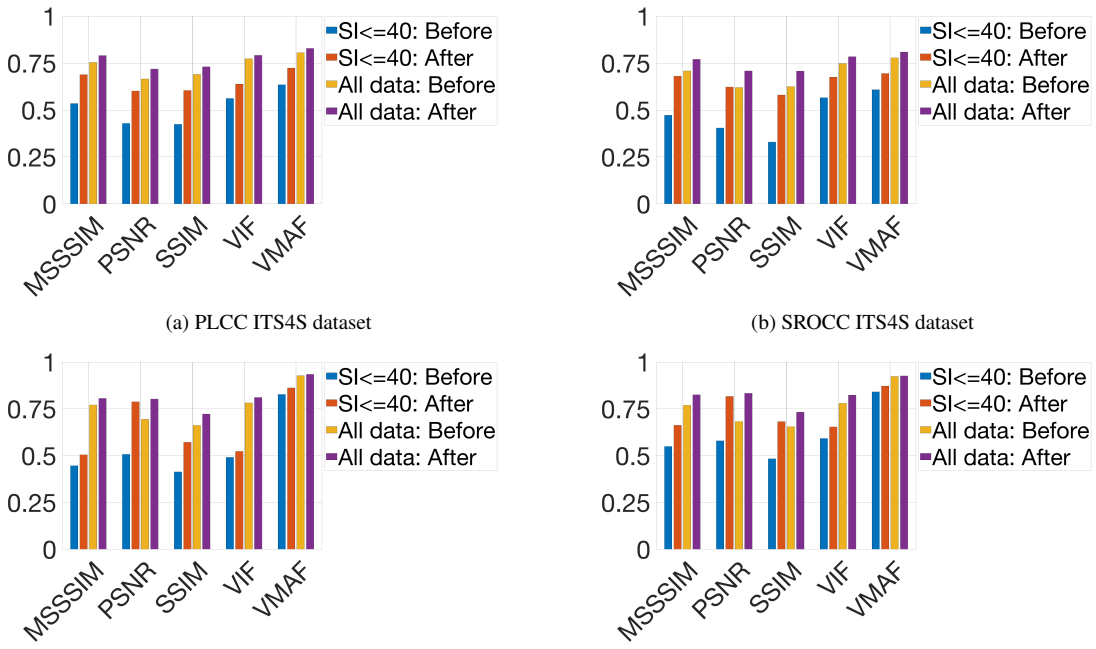
(a) PSNR



(b) VMAF

Figure 5. *VQMs tend to overestimate the perceived visual quality when used to assess the quality of sequences with low SI. Y axis represents the MOS estimation error.*

VQMs accuracy and comparing the performance of the VQMs considered in this study before and after adding the correction term $P_{VQM}(SI)$. As already mentioned we performed a polynomial fitting of each VQM score to the MOS. The coefficients of $f_{VQM}(VQM)$ as well as those of $P_{VQM}(SI)$ for each VQM have been computed using the least square method on the ITS4S dataset. Figure 5 reports, for instance, the behavior of $P_{PSNR}(SI)$ and $P_{VMAF}(SI)$ as a function of SI, i.e., the correction term to be applied in case the PSNR and the VMAF are used to assess the perceived visual quality. Very similar graphs have been obtained for all the other VQMs considered in the study. Since the graph shows that $P_{VQM}(SI)$ is a negative and increasing function, recalling the equality in (3) we can claim that the VQMs tends to overestimate the perceived quality of sequences providing from



(a) PLCC ITS4S dataset

(b) SROCC ITS4S dataset

(c) PLCC Netflix dataset

(d) SROCC Netflix dataset

Figure 6. *The numerical experiments demonstrate the effectiveness of the proposal since higher correlation coefficients are observed after the correction term is added. Computed on: ITS4S dataset. Tested on: Netflix dataset.*

sources characterized by low SI and that the lower the SI the higher the overestimation tends to be.

Figure 5 shows the PLCC and the SROCC between the VQMs and the MOS before and after adding the correction term i.e., considering the influence of low SI. The results are presented separately for sequences whose source is characterized by low SI ($avgSIsrc \leq 40$) and for all the sequences. It can be observed that in all cases higher correlation coefficients are obtained after adding to the VQM estimation the correction term.

Figure 6 shows that, despite all the coefficients of $f_{VQM}(VQM)$ and $P_{VQM}(SI)$ have been determined by using data from the ITS4S dataset, the approach demonstrates its effectiveness also on the Netflix dataset that has never been used at any stage to estimate the model parameters.

Conclusion

In this work we relied on the ITS4S dataset to investigate the behavior of well-known video quality measures. First, we extended the dataset with well-known full-reference measures, as well as spatial and temporal features. Relying on a statistical tools we investigated the similarity between the categories of contents in the dataset finding the most related ones. This could be useful to define heterogeneous subsets of sequences useful for, e.g., machine learning algorithms that needs carefully constructed training and test sets. An in-depth analysis of the newly added dataset values showed that the SI index is strongly related to the accuracy of VQMs when predicting the subjective quality. More precisely, the study revealed that VQMs tend to overestimate the perceived visual quality of contents generated from sources with low SI. On the basis of this observation we designed an improved MOS prediction scheme. Such scheme has been calibrated on the ITS4S dataset then tested on the same dataset as well as on the Netflix public dataset, showing that the effectiveness the proposal goes beyond the original ITS4S dataset used for this study.

Acknowledgment

This work has been supported in part by the Politecnico di Torino Interdepartmental Center for Service Robotics (PIC4SeR) https://pic4ser.polito.it.

References

[1] Institute for Telecommunication Sciences (ITS), https://www.ntia.doc.gov/office/ITS (2019).

[2] S. Winkler and P. Mohandas, The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics, IEEE Transactions on Broadcasting, 54, 3, pp. 660-668 (2008).

[3] NTIA/ITS, Consumer Digital Video Library (CDVL), http://www.cdvl.org (2020).

[4] M. H. Pinson, ITS4S: A Video Quality Dataset with Four-Second Unrepeated Scenes, NTIA Technical Memo TM-18-532 (2018).

[5] ITU-T, Subjective video quality assessment methods for multimedia applications, ITU-T Rec. P.910 (2008).

[6] ITU-T, Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM), ITU-T Rec. J.149 (2004).

[7] P. Hanhart and R. Hahling, Video Quality Measurement Tool (VQMT), http://mmspg.epfl.ch/vqmt (2013).

[8] Netflix, VMAF - Video Multi-Method Assessment Fusion, https://github.com/Netflix/vmaf (2019).

[9] P. Lebreton and W. Robitza and S. Göring, SITI, https://github.com/Telecommunication-Telemedia-Assessment/SITI (2019).

[10] G. Michael, Correspondence analysis in practice (3rd ed.), CRC Press, 2016.

[11] R version 3.3.3, https://www.r-project.org/ (2019).

Author Biography

Lohic Fotio Tiotsop received his MSc degree in Mathematical Engineering from Politecnico di Torino, Italy. Since 2018 he is a Ph.D. student in Control and Computer Engineering at Politecnico di Torino. His primary research interests are advanced statistical methods and machine learning algorithms applied to multimedia problems.

Antonio Servetti received his PhD degree in Computer Engineering in 2004 at the Politecnico di Torino, where he is currently assistant professor. His research focuses on web based multimedia communications and speech/audio processing.

Enrico Masala received his PhD degree in Computer Engineering in 2004 at the Politecnico di Torino, where he is currently associate professor. His main research interests include multimedia quality optimization of communications over packet networks, with special attention to particular scenarios such as remote control applications, 3D video, cloud for multimedia. He is also involved in the management of the Politecnico di Torino Interdepartmental Center for Service Robotics (PIC4SeR).

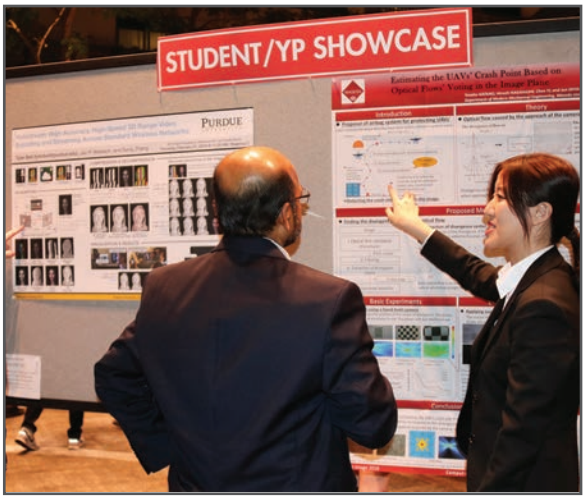
JOIN US AT THE NEXT EI!

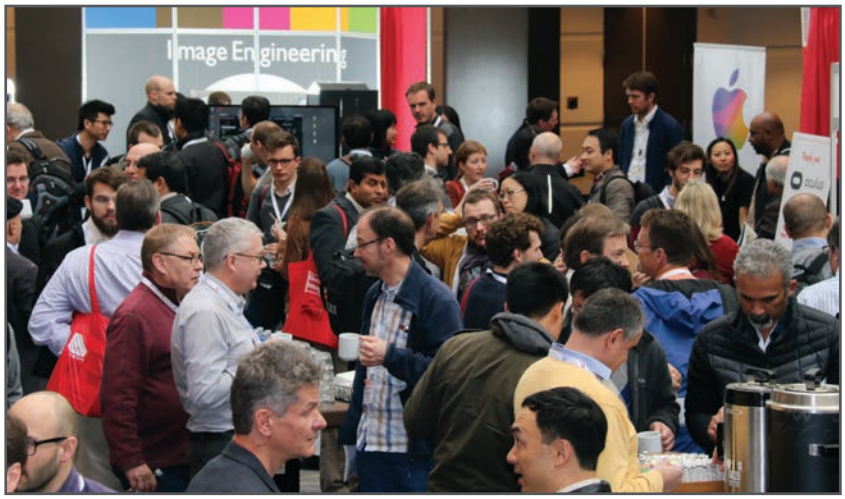
IS&T International Symposium on

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