https://doi.org/10.2352/ISSN.2470-1173.2020.11.HVEI-092
This is a work of the U.S. Government.

Open Software Framework for Collaborative Development of No Reference Image and Video Quality Metrics

Margaret H. Pinson; NTIA/ITS; Boulder, CO, USA
Philip J. Corriveau; Pacific University College of Optometry; Forest Grove, OR, USA
Mikołaj Leszczuk; AGH University of Science and Technology; Krakow, Poland
Michael Colligan; Spirent; San Jose, CA, USA

Abstract

This paper describes ongoing work within the video quality experts group (VQEG) to develop no-reference (NR) audiovisual video quality analysis (VQA) metrics. VQEG provides an open forum that encourages knowledge sharing and collaboration. The VQEG no-reference Metric (NORM) group's goal is to develop open-source NR-VQA metrics that meet industry requirements for scope, accuracy, and capability. This paper presents industry specifications from discussions at VQEG face-to-face meetings among industry, academic, and government participants. This paper also announces an open software framework for collaborative development of NR image quality Analysis (IQA) and VQA metrics <https://github.com/NTIA/NRMetricFramework>. This framework includes the support tools necessary to begin research and avoid common mistakes. VQEG's goal is to produce a series of NR-VQA metrics with progressively improving scope and accuracy. This work draws upon and enables IQA metric research, as both use the human visual system to analyze the quality of audiovisual media on modern displays. Readers are invited to participate.

Introduction

According to Cisco [1], "Globally, IP video traffic will be 82 percent of all IP traffic (both business and consumer) by 2022, up from 75 percent in 2017." Escalating video consumption drives the industry to seek more wireless bandwidth and higher visual quality at lower bandwidths. With the varied methods for content generation and distribution, better standalone tools are a must to drive experiences consumers expect. Improved methods to evaluate visual quality will help industry develop products and improve services. The missing component is no-reference (NR) metrics that perform image quality assessment (IQA), video quality assessment (VQA), and audiovisual quality assessment (AVQA).

Traditionally, the goal of IQA, VQA, and AVQA research is a single value that estimates the overall quality. From an industry standpoint, this is informative but not actionable. So, what if the quality is fair? To act, industry needs to know why the quality is bad and how to improve the quality. Most industry applications for NR metrics require root cause analysis (RCA). There have been NR-IQA tools developed from a camera capture perspective, but these tools do not take into account temporal changes or distribution concerns.

Another major problem is that IQA and VQA researchers often focus on impairments that diverge from industry applications. For example, IQA researchers are typically limited in scope to traditional impairments, such as JPEG compression, Gaussian blur, and white noise. Analyses indicate that NR-IQA and NR-VQA metrics developed for this narrow use case yield dramatically reduced performance when applied to the broad application of consumer content [2]. Products and services are starting to leverage visual processing algorithms and artificial intelligence (AI) based image manipulation to "enhance" quality (e.g., when upscaling for the target display). We don't have tools that address this use case, let alone the others that arise. Improved communication between industry and academia is needed to realize the vision of an NR-IQA or NR-VQA metric that industry can deploy in a broadcast or consumer workflow.

This paper is split into two main topics. First, we will summarize industry needs around NR-VQA metrics, based on discussions within the Video Quality Experts Group (VQEG). Second, we will present an open software framework for collaborative development of NR-IQA and NR-VQA metrics. This framework provides the tools and resources needed to conduct NR-IQA and NR-VQA research for the broad application of commercial content. By encouraging metric re-use, code sharing, and open data, open collaboration can produce robust solutions where private research and development has failed.

Industry Needs

NR-IQA and NR-VQA metrics are typically envisioned as real time substitutes for mean opinion scores (MOS) from subjective tests. However, the NR-VQA metric cannot simply estimate the mean opinion score ($\widehat{MOS}$) to predict overall quality. Decision makers need confidence intervals (CI) to understand whether the difference between two $\widehat{MOS}$ values is large enough to be significant. Subjective tests conducted on the absolute category rating (ACR) scale only have a CI of ≈0.5 on this [1..5] scale. In the absence of CIs, NR metric users assume infinite precision.

Video service providers need RCA to accurately identify specific quality problems (see Fig. 1). Professional video content is expensive to produce. Broadcasters treat footage carefully and their workflows include multiple quality checks. Two impactful checkpoints for NR-VQA metric deployment in broadcast workflows are ingesting real-time, on-location video streams (e.g., live sporting events) and ingesting third party video streams [3].

Internet service providers face similar challenges for user-generated content (UGC). YouTube ingests millions of user-generated videos every day, and quality analysis is important for compression and transcoding [4]. Traditionally, YouTube applies full-reference (FR) IQA metrics to each frame and aggregates (e.g., mean, or worst 5%). However, FR metrics require a pristine image or video to serve as the reference. FR-IQA fails when the uploaded video is non-pristine. FR-IQA cannot assess quality improvements. Simple aggregate statistics fail to model temporal changes to impairment levels. NR-VQA RCA enables video-ingress workflows that intelligently choose optimal sets of image filters and transcoding parameters for each video [4].

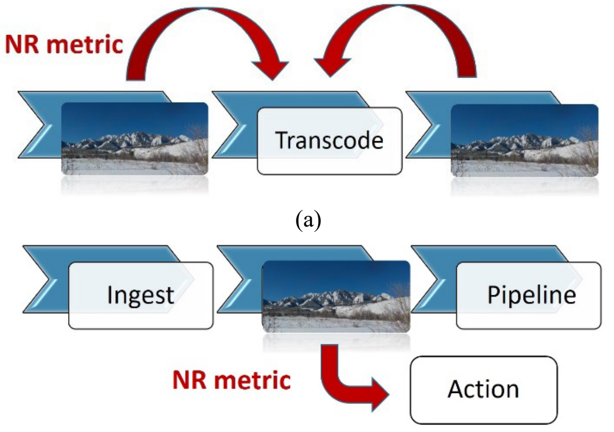

***Figure 1.** RCA could enable transcoder feedback loops (a) or detect problems that require intervention (b).*

Within broadcast workflows, management wants $\widehat{MOS}$. Workers need to detect underlying problems: misconfigured encoders, video upscaling, low quality de-interlacing, misconfigured cameras, missed videos, misconfigured video streams, and transmission medium errors that cause network traffic congestion or loss (e.g., atmospheric conditions during satellite transmission). These problems cause perceptible levels of macro-blockiness, blurriness, ringing, motion artifacts, black frames, noise frames, static video test signals, still video, and packet loss artifacts [5]. Audiovisual synchronization is a common problem for broadcasters, because the audio and video are often split and routed through different equipment. An RCA that identifies symptoms will help broadcasters locate underlying problems.

When industry uses $\widehat{MOS}$, they need a metric that can be easily modified to ignore some impairments and emphasize others, based on the user's task. Tailoring solutions to fit usage categories can be a huge advantage, especially if these implementations can learn or be trained. Broadcasters must reproduce impairments that reflect the producer's artistic intent, so $\widehat{MOS}$ may be misleading. Examples include noise and shaky camera work in found footage films like *The Blair Witch Project*, and the dark "Battle of Winterfell" episode of *Game of Thrones*. Broadcasters redefine $\widehat{MOS}$ to exclude quality losses from specific impairments, and they are not alone in this behavior. Codec developers redefine $\widehat{MOS}$ to ignore impairments associated with camera capture, aesthetics, scene content, and the camera operator's behavior. This viewpoint is pervasive among FR metric developers: $\widehat{MOS}$ is intrinsically limited to coding and transmission artifacts. Industry performs diverse tasks for different applications, and cannot be expected to retrain a machine-learning algorithm.

Internet video distribution workflows often use adaptive bit-rate (ABR) ladder encoding. The customer experience is directly impacted by the difference between the input video and the multiple video streams output by the ABR transcoder. Netflix uses VMAF, a FR-VQA metric, to improve ABR transcoding [6]. The video is segmented and encoded with diverse bit-rates, resolutions, etc. and evaluated with VMAF to find the optimal subset for ABR streaming.

NR-VQA metrics would enable an improved ABR transcoding workflow for impaired professional content, live environments [3], and consumer content. Ideally, the metric would hypothesize quality response curves for various encoding bit-rates and resolutions. Forecasted $\widehat{MOS}$s for conjectured encodings would shorten the development cycle.

When networks become congested, intelligent networks could use NR-VQA to consider the impact of bandwidth allocation decisions on user experience (UX). $\widehat{MOS}$ would suffice, but forecasted bit-rate/$\widehat{MOS}$ response curves would let the network's decisions be more "fair" from the perspective of human perception. NR-VQA metrics would let priority access protocols make better tradeoffs in response to the needs of priority and non-priority users for streaming video. An NR-VQA deployed in a network must be accurate for both professional and consumer generated content.

Video clients have analytic tools based on network parameters like bit-rate and buffering. The missing tool capability is an NR-VQA metric that measures client side video quality and returns that information to the provider. An NR-VQA metric on the client side would enable end-to-end quality ratings and could detect problems before the customer notices [3].

Camera and codec developers need NR-VQA metrics to optimize video encoding algorithms. Like broadcasters, codec developers need both $\widehat{MOS}$ and RCA. The NR-VQA metric must be extremely fast (real-time if possible), and it must understand quality impacts of the entire camera capture pipeline (i.e., sensor, image processing, encoder, decoder, and display).

Coding decisions are made independently on small blocks. This community needs the NR-VQA metric to scale down to 64 pixel × 64 pixel × 2 frames. Humans cannot evaluate the quality of 0.03 seconds of video without the surrounding context, so we must create training data that enables this extrapolation, as proposed by [7]. Camera capture is one of the few applications where the NR-VQA metric absolutely must be pixel based; bit-stream information does not yet exist. An NR-VQA metric would also help the camera optimize performance for applications with different quality needs.

First responders want intelligent cameras that understand how their needs differ from broadcasters and consumers. First responders use cameras for mission critical response in environments that stress cameras—inclement weather, smoke, dark nights, and jiggling camera mounts. Law enforcement officers need to meet the evidence needs of the courts to accurately portray situations and events. Forensic video analysts want the entirety of each video frame to be in focus, because individual frames will be extracted for use as photographs. During a snowstorm, a typical camera tries to reproduce the falling snow, but first responders watching video surveillance monitors want to see distant people or read license plates. First responders need an NR-IQA metric so that the camera can warn them about problems while there is still a chance to take another photograph [8].

Another application with unique needs is AI systems, such as autonomous vehicle systems and video analytics. Out-of-service, the NR-VQA metric would serve as a prefilter, detecting whether the quality is high enough for the AI algorithm to succeed. In-service, the NR-VQA metric would enable camera control feedback loops. The system could pan, zoom, re-focus, increase the bit-rate, turn on a light, or change camera feeds to boost the AI algorithm's accuracy [9]. Like codec developers, AI systems need scalable NR-VQAs that can operate on small regions of interest.

Metric Capabilities

Open source usage rights are critical for viable collaboration and widespread adoption. Licensing restrictions hinder collaboration and cause metrics to languish unused. Industry needs to understand the metrics and to trust their analyses ($\widehat{MOS}$ and

RCA). Researchers need industry to provide feedback. Industry feedback contributes to the pool of knowledge and helps researchers focus on high-impact problems. Particularly welcome are sets of application-specific images or videos that depict a specific impairment. Researchers need these datasets to train RCA algorithms.

Our goal is a single metric that predicts $\widehat{MOS}$ and RCA for both IQA and VQA. The importance of RCA was demonstrated in the previous section. Conventionally, IQA and VQA are separate lines of research. Today, the displays are identical, and the cameras include common electronics. We will gain increased understanding by merging IQA, VQA, and eventually AVQA into a single line of research.

Our training data must include quality problems associated with the real world subject, the camera operator's actions, aesthetics, and the entire camera capture pipeline. The metric must understand and accommodate these quality problems, even if $\widehat{MOS}$ ignores them. Users will apply the metric to new content, and the metric must have a minimal loss of accuracy. Users will also apply the metric to out-of-scope impairments, and performance must degrade gracefully. Reduced accuracy is acceptable; random results are not.

The metric must be fast, to enable real-time implementation on video streams. An NR-IQA that takes minutes for a single image is impractical for NR-VQA analysis. The metric must run on any resolution or frame rate.

To achieve this lofty goal, we accept two limitations. First, quality estimates assume the media is scaled for a particular display (e.g., results are reported for a 1920 × 1080 display). Thus, we ignore the complex question of how to evaluate the added value of 40 megapixel (MP) image over a 5 MP image, when both are viewed on a (1920 × 1080) monitor. Second, we evaluate the immediate quality response, using very short videos without scene cuts. The motion picture experts group (MPEG) limits content in this way when evaluating proposed coding algorithms. Work in the International Telecommunication Union (ITU) Study Group 12 demonstrates that temporal integration of quality fluctuations can be studied separately and applied as post-processing.

Software Framework

Literature identifies a core of innovative NR-IQA and NR-VQA metrics. Feedback from industry indicates none of the available metrics meets their needs around scope, accuracy, and features. Common problems include ambiguous licensing terms, unavailable source code, slow run speed, insufficient training data, failure to provide RCA, and exaggerated metric performance. Another major concern is that NR metrics are a black box that cannot be understood, and thus cannot be trusted.

Where individual efforts have been unsatisfactory, success may be possible by pooling industry, academic, and government resources in open collaboration. To that end, VQEG formed the no-reference metric (NORM) working group. Our goal is to gather knowledge of industry requirements, produce datasets of topical images and videos, create software tools, and establish a series of metrics.

This section announces a GitHub repository that contains an open source software framework for collaborative development of NR-IQA and NR-VQA metrics [10]. The initial version of this software framework was provided by NTIA/ITS, with the intention that all interested parties will contribute to a growing body of code. This framework provides:

- Open source license
- List of training datasets
- Data structure to codify datasets
- Standard function interface for metrics
- Control software, to compute metrics on multiple datasets
- Analysis tools

NR-IQA and NR-VQA metrics are typically trained using three or fewer datasets, most likely due to limited computation power, difficulty obtaining datasets, and logistics. Storage and computation problems are an inevitable byproduct of video research. The GitHub repository resolves the other two problems by identifying suitable training databases and a software framework that provides logistic support for handling thousands of images and videos from diverse datasets.

Most publicly available subjective datasets are not ideal for training NR-IQA or NR-VQA metrics. Publicly available datasets are far too small for machine learning. Traditional IQA and VQA experiments use a small set of pristine source media, which does not address the need for robust response to new content. Traditional VQA experiments use 8+ second videos with temporal changes, which does not adhere to our "immediate quality response" goal. The impairments may be outdated or unrelated to industry use cases. The GitHub repository mostly identifies newer experiments, and more training data is needed.

The software framework establishes a data structure that describes a subjective dataset (e.g., file names, subjective ratings, resolution displayed to subjects). This data structure logically divides the media (images or videos) into categories specified by the experiment design. For example, the CCRIQ dataset has categories for display on a 4K monitor (2160 × 3840) and display on an HD monitor (1080 × 1920) [11]. The GitHub repository contains pre-filled structures plus functions that create structures for new datasets. These MATLAB® functions[1] are named *import_dataset* and *export_dataset*.

One category established for all datasets is *training* vs *verification*. Of the media in each dataset, 90% are available for training and 10% are set aside for metric verification. The verification data are only used to report how performance drops on data that was *never* used for training. This addresses industry's need for unexaggerated performance evaluations—provided there is minimal overlap of scenes and systems between the *training* and *verification* categories. Note that machine learning must further split the 90% *training* data into subsets for training & validation iteration.

The software framework establishes a standard interface for calculating NR features, parameters, and metrics. This standard interface is referred to as an NR feature function (NRFF). Function *calculate_NRpars* does the heavy lifting of running the NRFF on multiple datasets. Function *calculate_NRpars* provides code to read media (images or videos), split videos into frames, perform color

[1] Certain commercial equipment, materials, and/or programs are identified in this article to specify adequately the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the program or equipment identified is necessarily the best available for this application.

space conversions, rescale, deinterlace, parallel process, etc. Basically, *calculate_NRpars* provides logistic support to calculate NR metrics on diverse datasets.

NR features hold intermediate calculations (e.g., local estimate of blurring or noise). The NRFF produces a number, vector, or matrix of values for each image or video frame. For example, the NR feature may divide each media into ≈100 regions of roughly the same size using function *divide_100_bocks.m* and apply a calculation to each region.

NR features naturally fall into four types. Spatial impairment (SI) features operate on individual images or video frames, as per NR-IQA metrics. Temporal impairments (TI) features operate on two subsequent frames, to analyze motion. Images are converted into still video before calculating TI features. Less common are features that operate on an entire video at once, for example to perform a 3D Fourier Transform. The fourth type of NR feature manipulates bit-stream information. An NRFF interface that is planned but not yet available will provide motion vector and quantization protocol (QP) information from video bit-streams. Purists would bar bit-streams from NR research but, realistically, bit-streams are available for most industry applications.

Where NR metrics predict overall quality (typically $\widehat{MOS}$), NR parameters also have a single value for each medium but serve as an intermediate result between NR features and NR metrics. The framework assumes a workflow where the researcher chooses an impairment; designs several NR features and parameters; calculates the NR features and parameters; analyzes the results; iterates until satisfied; and ultimately combines NR parameters into an NR metric using linear regression. NR features are saved, so that the researcher can quickly try many different ideas to calculate NR parameters from NR features. This workflow assumes that each NR parameter focuses on a specific impairment and that machine learning (if used) is conducted on NR parameters (to yield RCA) and not on NR metrics (to yield $\widehat{MOS}$).

This workflow addresses three industry concerns. First, the NR metric naturally provides RCA via the NR parameters. Second, $\widehat{MOS}$ can be easily modified to remove an impairment the user wants to ignore (i.e., by removing an NR parameter from the final equation). Third, the metric is less of a black box. Linear regression produces easily understood equations, and the motivation of each RCA can be understood even if the algorithm is incomprehensible.

The GitHub repository provides several tools for analysis and metric building. Function *analyze_NRpars* calculates statistics and creates scatter plots to help the user analyze parameters. This function is part of the "iterate until satisfied" step of the workflow. Function *compromise_NRpars* calculates statistics and creates plots to help the user understand whether two NR parameters complement each other. Function *export_NRpars* saves NR parameter data to a spreadsheet, so that other programs can be used to train metrics.

Root Cause Analysis

RCA provides a more difficult challenge than $\widehat{MOS}$. RCA can describe the impairments that contribute to a subjective score (see Fig. 2); or RCA can identify a transformative or distributive process that creates impairments (e.g., compression, scaling, transmission medium errors, camera, or display). The former definition provides a pragmatic starting point for NR metric development but ultimately industry wants the latter (cause identification). In this section, we describe five strategies for training RCA.

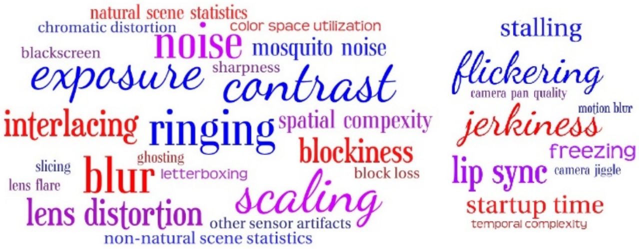

Figure 2. VQEG identified 23 spatial artifacts (left) and 10 temporal artifacts (right) that influence root cause analysis.

The first strategy for RCA research is to remove the influence impairments *other* than the one being studied can explain, and to remove it also from the MOSs. This increases the value of datasets like CCRIQ where MOS is influenced by a variety of confounding impairments. When researching noise, for example, the researcher could gather a set of NR parameters for impairments other than noise, create a metric, calculate residuals ($MOS - \widehat{MOS}$), and then evaluate potential noise NR parameters against those residuals. The GitHub repository enables this strategy by encouraging RCA metric sharing and the NRFF interface.

Appendix C of [12] proposes a second strategy for media with confounding impairments: perform a subjective test where subjects rate the influence of several RCA factors on each media's quality. The time and expense of the subjective test would increase. No such subjective data is currently available.

The opposite solution is a *challenge dataset*—a set of images or videos that demonstrate a single impairment, while avoiding others. Challenge datasets must include high, medium, and low levels of the impairment, plus unimpaired media. While other impairments cannot be fully eliminated, their influence must be minimized. The *its4s4* dataset [13] demonstrates the challenge dataset concept for camera pan impairments, and function *nrff_IPSpan.m* contains the resulting NR-VQA parameter. Challenge datasets can use the traditional ACR scale, which simplifies subjective testing and algorithm development.

The third strategy is to create a challenge dataset as a field study. We propose this strategy as an impactful way for industry to encourage RCA metric development—show us what you want. The field study emphasizes realism and scene variety to demonstrate the authentic workflow and diverse response of a real application. We recommend at least 100 media, either images or 4 s videos without scene cuts. Unrepeated scene experiment designs [14] are preferred, so the RCA has a robust response to new content (e.g., each video depicts a different scene). Industry involvement would be limited to selecting media; researchers can then perform subjective tests and tackle algorithm development.

The fourth strategy is to create a challenge dataset as a lab study. These challenge datasets contain a full matrix of scenes and impairments. This strategy allows researchers to scrutinize the RCA algorithm's biases for different scenes. The disadvantage is reduced realism (e.g., simulated impairments, limited subject matter variety). Leszczuk et al. [15] demonstrate this strategy, illustrated through four challenge datasets and seven RCA algorithms (e.g., exposure time, noise, and freezing). These RCA algorithms, referred to as key performance indicators, detect the presence of an impairment, measured as a Boolean.

The fifth strategy is to take advantage of experts. Fernández and Leszczuk [16] demonstrate this approach for audio-to-video synchronization (AV sync or lip-sync). The authors based the measurement of AV sync distortion on two components: (i) the degree of mouth opening and (ii) the presence of speech. These are measured successively through modules called lip movement tracker and voice activity detector. They obtained video samples with perfect AV sync to create training data (ground truth data) for each video with "0 delay" (meaning with no AV sync error). Oher delays were generated artificially.

Datasets that portray the complex impairment interactions of a real application are valuable for double-checking RCA algorithms. For example, the CCRIQ dataset [11] depicts the camera capture problems of 23 cameras. A drop in performance is expected (due to the presence of other impairments) but over-trained RCA metrics will exhibit poor behavior that can be seen on scatter plots (e.g., as a random scattering of data points). General datasets can indicate whether the proposed RCA metric detects the unintended impairment. One approach is to compute a linear fit between $\widehat{MOS}$ and several RCA metrics that track different impairments. For example, let us consider the CCRIQ dataset (which contains blur and noise impairments), a proven RCA metric that detects blur, and a new RCA metric that detects noise. If the noise metric's contribution to the linear equation is not statistically significant, then the "noise metric" may instead be detecting sharp edges.

Challenge datasets can be made freely available on the Consumer Digital Video Library (CDVL, www.cdvl.org). CDVL videos can be used for this purpose, provided the modified videos are redistributed on CDVL.

Conclusion

The Industry at large needs to establish an open source work-stream to develop, refine, validate, and deploy NR-IQA and NR-VQA tools. The framework in this paper supports a proposal that would elevate the quality analysis across several use cases that are acceptable to both professionals and consumers. The framework is available at https://github.com/NTIA/NRMetricFramework

Acknowledgements

This report was partially supported by the Polish Ministry of Science and Higher Education with the subvention funds of the Faculty of Computer Science, Electronics and Telecommunications of AGH University.

References

[1] Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper.

[2] M. Pinson, *Analysis of No-Reference Metrics for Image & Video Quality Consumer Applications,* NTIA Memorandum TM-20-546, Jan. 2020.

[3] J. Webb, "Broadcast and Creating Use Case for Non-Reference," *Madrid VQEG Meeting*, VQEG_NORM_2018_003, Mar. 2018.

[4] Y. Wang, "Quality Analysis for UGC Videos," *Mountain View VQEG Meeting*, VQEG_NORM_2018_121, Nov. 2018.

[5] M. Pinson, "Broadcast Use Case Discussion," *Madrid VQEG Meeting*, VQEG_NORM_2018_022, Mar. 2018.

[6] Z. Li et al., "VMAF: The Journey Continues," *The Netflix Tech Blog*, Oct. 25, 2018.

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

[8] M. H. Pinson, "Technology Gaps in First Responder Cameras," NTIA Technical Memo TM-17-524, May 2017.

[9] M. Pinson, "First Responder Use Case," *Madrid VQEG Meeting*, VQEG_NORM_2018_021, Mar. 2018.

[10] National Telecommunications and Information Administration, Institute for Telecommunication Sciences, "NR Metric Framework," https://github.com/NTIA/NRMetricFramework, accessed 2/5/2020.

[11] M. A. Saad et al., "Impact of Camera Pixel Count and Monitor Resolution Perceptual Image Quality," *Colour and Visual Computing Symposium (CVCS)*, 2015, Gjovik, Norway, 25-26 Aug. 2015.

[12] S. Wolf and M/ H. Pinson, "Video Quality Measurement Techniques," NTIA Technical Report TR-02-392, June 2002.

[13] M. Pinson and S. Elting, "ITS4S4: A Video Quality Study of Camera Pans," NTIA Technical Memo TM-20-545, Dec 2019.

[14] L. Janowksi, L. Malfait, and M. H. Pinson, "Evaluating experiment design with unrepeated scenes for video quality subjective assessment," *Quality and User Experience*, Dec. 2019.

[15] M. Leszczuk, M. Hanusiak, M. C. Q. Farias, E Wyckens, G Heston, "Recent Developments in Visual Quality Monitoring by Key Performance Indicators," *Multimedia Tools and Applications* 75 (17), 10745-10767, 2016.

[16] I. B. Fernández, M. Leszczuk, "Monitoring of Audio-Visual Quality by Key Indicators," *Multimedia Tools and Applications* 77 (2), 2823-2848, 2018.

Author Biography

Margaret H. Pinson received her BS and MS in computer science from the University of Colorado at Boulder (1988 & 1990). Since then she has worked for the Institute for Telecommunication Sciences (ITS) in Boulder, CO, where she investigates improved methods for assessing video quality. Technically, she is known for her objective video quality metrics, for her support of video quality metric validation efforts, and for research to improve subjective video quality test methods.

Philip Corriveau received his Bachelors of Science in Psychology from Carleton University. He has been a user experience researcher for over 30 years specializing in quantitative metrics. Philip is currently a Senior Principal User Experience Researcher at the Vision Performance Institute and Adjunct Assistant Professor, Pacific University College of Optometry.

Mikołaj Leszczuk received his BSc/MSc in telecommunications from the AGH University of Science and Technology (2000), his PhD in telecommunications from AGH University of Science and Technology (2006) and his DSc in information and communication technology from AGH University of Science and Technology (2017). Since then he has worked as Associate Professor in the AGH Faculty of Computer Science, Electronics and Telecommunications in Kraków, Poland. His work has focused on the development of audio-visual quality evaluation systems. He is on the Board of VQEG.

Michael Colligan received his BS in Computer Science from the University of Nebraska and his MS in Software Engineering from San Jose State University. He has been a researcher in video compression, distribution, and test for over 25 years. Michael is currently the Video Systems Architect for Spirent Communications.