NTIA Technical Memorandum 20-547

Analysis of No-Reference Metrics for Image and Video Quality of Consumer Applications

Margaret H. Pinson

technical memorandum

U.S. DEPARTMENT OF COMMERCE • National Telecommunications and Information Administration

NTIA Report 20-547

Analysis of No-Reference Metrics for Image and Video Quality of Consumer Applications

Margaret H. Pinson

U.S. DEPARTMENT OF COMMERCE

January 2020

DISCLAIMER

Certain products, technologies, and corporations are mentioned in this report to describe image and video technologies. The mention of such entities should not be construed as any endorsement, approval, recommendation, prediction of success, or that they are in any way superior to or more noteworthy than similar entities that were not mentioned.

CONTENTS

FIGURES

TABLES

GLOSSARY OF TERMS

NR Feature Intermediate calculation while calculating NR Metrics (e.g., local estimate of blurring or noise)

NR Metric A metric that predicts the quality of an image or video using only the image or video itself (i.e., pixels) without referencing the bit-streams, coding parameters, or a higher quality version of the image or video

NR Parameter An NR metric that measure the quality response of a single impairment, used to provide root cause analysis

Root Cause Analysis NR metric analysis that divides the overall quality into multiple factors, to explain why the quality dropped

ANALYSIS OF NO-REFERENCE METRICS FOR IMAGE AND VIDEO QUALITY OF CONSUMER APPLICATIONS

Margaret H. Pinson[1]

This paper analyzes the performance of eight no-reference (NR) metrics. Seven assess image quality (BRISQUE, CurveletQA, IL-NIQE, NIQE, OG-IQA, QAC and SSEQ) and one assesses video quality (VIIDEO). The challenge we address in this paper is moving from research silos to a broadly applicable metric. Our analyses use six new subjective datasets that characterize modern cameras and high performing networks. Five datasets were designed around consumer applications and no-reference metric development. The sixth dataset was designed for full-reference metric analyses; we present a technique to modify older datasets for NR metric development. Our analyses show a need for more research and development. The NR metrics were inaccurate for consumer applications.

Keywords: BRISQUE, CurveletQA, IL-NIQE, image quality, NIQE, no-reference metrics, NR, OG-IQA, QAC, SSEQ, video quality, VIIDEO

1. INTRODUCTION

Important industrial decisions often rely on ad hoc quality evaluations instead of image quality assessment (IQA) and video quality assessment (VQA) algorithms. Many industrial workflows can only accommodate no-reference (NR) metrics—a “blind” metric that uses an image or video to predict its own quality. Ad hoc evaluations of NR metrics by private sector video quality professionals (from private communications) indicate that the available metrics are too inaccurate to be trusted.

The better-known metrics, like *NIQE* [1] and *BRISQUE* [2], were trained on datasets like the *LIVE Image Quality Assessment Database* [3] that use a full matrix of source scenes (SRCs) and impairments (HRCs, hypothetical reference circuits). Databases designed for NR metric development use unrepeated scene experiment designs—each image or video is included only once instead of repeatedly [4]—and include camera impairments. This better represents the media encountered by consumer applications.

Unlike full reference metrics, NR metrics must understand quality problems caused by the camera, the camera operator, and interactions between the camera and the camera operator. Our target audience (a typical consumer) cannot differentiate between impairments caused by the codec, the camera capture, the video format, and the camera operator. We must expect this behavior from our NR metric’s users. Even if the goal is an NR metric that only predicts coding

[1] The author is with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.

impairment quality, the metric must understand, evaluate, and explain other factors—or people will not trust its predictions.

Were such a metric to exist, image and video quality assessment could be more than just an afterthought. A crime scene investigator's camera could warn the user that an image will not meet evidentiary needs, while there is still time to take another picture. A surveillance camera could automatically adjust its settings to meet the needs of a video analytic running in the cloud. An intelligent network could analyze the quality of various video streams and make intelligent decisions on how to apportion limited bandwidth. A content delivery provider could analyze user-generated content, detect problems and choose optimal coding parameters.

This paper examines NR-IQA and NR-VQA metrics from the point of view of industrial and consumer needs. We will clarify our application, specify subjective datasets suitable for training NR metrics, and evaluate the performance of open source NR metrics. Our goal is to demonstrate proposed directions for needed research into NR metrics for consumer applications.

This paper does not provide a fair validation of the metrics for their intended use case. We take freely available metrics with narrow scopes and evaluate their performance on the broad application of consumer content. Our analysis includes camera impairments, compression artifacts, both professional and amateur camera operators, and a wide variety of source material. Conversely, the NR metrics were trained on traditional data, which emphasized JPEG compression, white noise, Gaussian blurring, and professional camera operators. Our goal is to understand how their performance degrades in response to unforeseen data.

2. CONSUMER APPLICATION

This paper focuses on the range of quality that consumers expect from modern cameras combined with high performing networks. Applicable services include professional and amateur photography, broadcast television, adaptive video streaming, video surveillance, videoconferencing, and websites. We assume carrier-grade network performance with high throughput, very low latency, and low bit-error rate.

We do not differentiate between images and video. Conventionally, IQA and VQA are considered separate lines of research. Media services present a mixture of videos, paused videos, and images on a single display. VQAs must be able to assess images.

Our assessment must consider the quality impact of the camera, the actions of the camera operator, the scene depicted, and state-of-the-art codecs operating at bitrates suitable for modern applications. Our target audience (a typical consumer) cannot differentiate between impairments caused by the codec, the camera capture, the image or video format, and the camera operator. Even if the goal is an NR metric that only predicts coding impairment quality, the metric must understand, evaluate, and explain other factors—or people will not trust its predictions.

Our goal is to extrapolate the quality of two frames of video, even though people cannot perceive individual frames in isolation (e.g., an image displayed for 0.03 seconds). Video encoders contain complex networks of algorithm choices. Basically, the encoder partitions the video into small sub-regions (e.g., 64 pixels × 64 pixels × 2 frames) and chooses encoding options for each sub-region. Codec developers want insights into the quality impact of sub-regions on the overall video quality. This would lead to coding efficiency improvements. Thus, video sequences are limited to the shortest duration that our subjects could comfortably rate, which is 4 s [5].

Our application excludes confounding issues and impairments that will become less common over the next decade. Examples include transmission errors, temporal integration, legacy systems, low bit-rate transmission, and network congestion. Temporal integration (i.e., dynamic changes over time) will be studied separately. ITU-T Recommendation P.1203 demonstrates that temporal integration can be studied separately and applied as post-processing (e.g., estimate the quality of a 5 minute video by dividing the video into 1 second segments and integrating these individual quality estimations).

3. DATASETS FOR NR METRIC RESEARCH

Our analysis is only as good as our data, so we must first ensure the quality of our validation data. This section describes the unseen data used to analyze NR metrics. These summaries are provided to help the reader understand the strengths and weaknesses of each dataset. All subjective tests are naturally limited in scope and thus incomplete, which in turn limits the accuracy of our analyses in Section 5.

We will use six datasets to evaluate NR metrics. Most of these datasets are freely available for research and development purposes (see [6] and [7]). The rest will be made available soon. Unless otherwise specified, all datasets use the Absolute Category Rating (ACR) scale, implemented with five levels.

Our data must exercise the main variables that impact the NR metric's performance: camera, codec, scene, video format, bit-rate, etc. The list of popular image and video codecs is short—JPEG, H.264/AVC, H.265/HEVC and WM9—and scene content is infinite in diversity. Thus, we need our training data to include a large number of scenes, plus impairments that span the full scope of our application.

Most freely available IQA and VQA datasets are ill-suited for NR metric training. They tend to include a small number of high quality source stimuli, few camera impairments and a large variety of coding and network impairments. Those datasets will tend to produce NR metrics that perform poorly on unseen source content and camera impairments (e.g., hand jiggle, too-fast pan, noise from poor lighting).

ITS4S [5] presents a simplified model of an adaptive video streaming service played to a 720p monitor (720 × 1280, 24fps). *ITS4S* uses an unrepeated scene experiment design (i.e., each sequence is unique). *ITS4S* contains 813 unrepeated scenes of 4 s duration, edited in sets from professionally produced, contribution quality footage (e.g., 34 segments from Netflix "El Fuente"). The raw footage was format converted to 720p 24fps, and the SRCs contain no further processing. The HRCs span H.264 bitrates from 2.34 Mbps to 0.512 Mbps and use a simplified bit-rate/resolution ladder. The encoding resolution drops linearly from (1280 × 720) at 2.34 Mbps to (512 × 288) at 512 Kbps. The videos were up-sampled to (1280 × 720) for the subjective test. The NR metrics are given these up-sampled videos. *ITS4S* includes footage with poor aesthetics, camera problems, and videography problems. *ITS4S* contains a few 720p 60fps SRC, which were eliminated from this assessment. If we ignore outliers, the *ITS4S* SRCs scores range from 3 to 5 (fair to excellent). Thus, the *ITS4S* dataset includes both coding artifacts and camera impairments.

ITS4S2 [8] contains 1,473 images from a variety of consumer cameras. *ITS4S2* provides diverse images to train NR metrics that span two tasks: entertainment and public safety. The *ITS4S2* dataset characterizes the entire camera capture pipeline: sensor, image processing encoder, decoder, and display. Most of the photographers were amateurs. Images are organized into themed sessions (e.g., landscapes, disasters). The images were obtained from Flickr® as follows:

- 29% from the *VIME Image Database* [9]
- 40% from public safety practitioners or similar content

- 31% chosen for specific subject shapes or topics (e.g., fireworks, parquet floors, sunsets, and blurred backgrounds)

The *VIME Image Database* was created by VQEG's video and image models for consumer content evaluation group (VIME), to support NR metric research development. *ITS4S2* contains mainly unique photographs, plus small sets of related images (e.g., an image with and without post-processing, two different camera settings). The *ITS4S2* analyses are based on 16 subjects using HD monitors (1920 × 1080).

CCRIQ [10]-[11] includes the same 16 scenes, photographed with 23 different cameras. The goal was to better understand the relationships among camera type, image pixel count, monitor resolution (HD vs 4K), camera characteristics (optics and post processing), and the overall perceived quality. Subjects rated each of the 392 images twice: once on an HD monitor (1920 × 1080) and once on a 4K monitor (3840 × 2160). The images range from 1 MP to 18 MP. The *CCRIQ* images were down sampled to both display resolutions before running metrics.

CCRIQ2 and *VIME1* are two sessions in the same experiment and will be analyzed jointly with the abbreviation *C&V* [12]. *CCRIQ2* contains 92 photographs that were photographed for the *CCRIQ* experiment but excluded due to size constraints. Thus, these photographs use the same 23 cameras as CCRIQ but 4 new scenes. *VIME1* contains 102 images from the *VIME Image Database* that do not appear in *ITS4S2*. Like the *CCRIQ* and *CCRIQ2* datasets, *VIME1* contains seven scenes photographed with multiple consumer cameras. The images were down-sampled to the display resolution (1440 × 900) before running metrics.

LIVE Public-Domain Subjective in the Wild Image Quality Challenge Database (Wild) [13] contains 1,162 photographs from diverse mobile devices. The dataset contains diverse scenes and camera capture impairments. Each image is 500 × 500 pixels. *Wild* uses a continuous, 100-level scale. To simplify this presentation, the mean opinion scores (MOS) were linearly mapped to a five-level scale.

Four of these five datasets contain only camera capture impairments. The *ITS4S* dataset could attribute undue accuracy to NR metrics that detect blurring, due to the linear relationship between resolution and bitrate. We need another dataset that contains coding impairments for a large variety of SRCs.

The best available option is the VQEG HDTV validation data [4]. VQEG provides MOSs that join the six datasets into a single dataset using overlapping video sequences. This is referred to as "superset MOS" data. Each dataset was designed around 10 s SRCs from 1080i 59.97fps, 1080i 50fps, 1080p 30fps, or 1080p 25fps footage. After eliminating transmission error HRCs and the overlapping sequences, the combined dataset contains 54 SRCs and 49 HRCs that span MPEG2 from 4 to 15 Mbits and H.264 from 1.5 to 13.5 Mbits. If we ignore outliers, the SRC scores range from 4 to 5 (good to excellent).

However, the VQEG HDTV SRCs contain changes in scene content and coding complexity that are outside the scope of our application. To address this problem, we create a faux dataset, *vqegHDcuts*. Each SRC was cut whenever the content or camera motion changed (e.g., at a scene cut, before and after a fade, before and after a camera pan). The MOS of the entire sequence was assigned to each segment, which adds error to the MOSs. This is an unprecedented technique, so the magnitude of this error is not known. The 54 SRCs were cut into 230 segments, each with similar content and similar amounts of motion throughout. The edit points for each HRC were corrected manually, to compensate for temporal warping. The *vqegHDcuts* dataset contains 2,145 video sequences.

To understand the overall response of NR metrics to camera capture, we will define *cam90%* to be 90% of the SRC images and videos in each of the first five datasets, chosen at random. The *vqegHDcuts* SRC are excluded, due to the unknown error stemming from the faux dataset creation process. *cam90%* includes a large variety of scenes and camera capture quality responses. The remaining 10% of original media are held in reserve as unseen data for future research. The MOSs from the five experiments are combined, without mapping, to remove lab-to-lab and experiment-to-experiment differences. *cam90%* contains 93% images, 7% videos and few coding artifacts.

We know this adds error to the *cam90%* MOSs. Mapping solutions require either overlapping sequences, which we lack, or objective metrics, which would bias our metric analyses. Moreover, we cannot estimate this error without a technique to map the datasets. From the point of view of psychology, the datasets are five realizations of the same question around consumer opinion of the quality of camera capture—but the large differences in experiment designs are problematic. Like the *vqegHDcuts* dataset, the *cam90%* analysis has an unknown error magnitude.

To understand NR metric response to codec bitrate, the *ITS4S* dataset is split into subsets by HRC. These analyses omit the same 10% of *ITS4S* SRC videos as *cam90%.*

4. NR METRICS AND NR PARAMETERS

We will analyze eight NR metrics published from 2011 to 2016 (see Table 1). Source code for these metrics is freely available, and the authors provide usage rights that include redistribution and commercial purposes. We omit older metrics (newer metrics are of more interest), unduly slow code, metrics that must be purchased, code that cannot be called as a subroutine, and metrics available only as a publication. Note that MATLAB®'s BRISQUE and NIQE code was used instead of the code provided by the metrics' authors, because we expect people are more likely to use the code provided by MATLAB. This MATLAB code occasionally returned not a number (nan), which was replaced by zero (0). None of the NR metrics were trained on the datasets in Section 3. See the Glossary of Terms for our definitions of NR feature, NR parameter, and NR metric.

The *Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)* [2] is an IQA that uses natural scene statistics and supports vector regression (a type of machine learning). MATLAB's *BRISQUE* code was used instead of the code provided by the University of Texas at Austin. When *BRISQUE* produced not a number (nan), zero was used instead.

The *Naturalness Image Quality Evaluator (NIQE)* [1] is an IQA that further develops the natural scene statistics concept presented in *BRISQUE*. *NIQE* bases the final quality prediction on maximum likelihood estimation and features extracted from training images. As with BRISQUE, MATLAB's *NIQE* code was used instead of the code provided by the University of Texas at Austin, and nan replaced with zero.

Quality-Aware Clustering (QAC) [14] is an IQA that is trained on a faux dataset. Pristine images were impaired with noise, blur, and compression artifacts. Subjective scores were created by apportioning FR-IQA ratings to different patches of the image. *QAC* builds on an earlier metric from the University of Texas at Austin.

SSEQ [15] is an IQA that uses local spatial spectral entropy features with a support vector machine.

CurveletQA [16] is set of 12 IQA parameters designed around the curvelet transform. Liu et al. [16] demonstrate the value of these parameters within an NR metric, rather than providing a particular NR metric. Their iterative analysis procedure involves dividing datasets randomly into training and testing subsets. By consequence, we will analyze the *CurveletQA* parameters, which are named *f1* through *f12*.

Integrated Local NIQE (IL-NIQE) [17] is an IQA that expands upon the earlier *NIQE*.

The *Oriented Gradients Image Quality Assessment (OG-IQA)* [18] is an IQA that contains six parameters designed around the local gradient orientation and magnitude, which are combined using AdaBoosting back-propagation neural network. The *OG-IQA* software makes six parameters available: gradient magnitude (*GM*), relative gradient orientation (*RO*), and relative gradient magnitude (*RM*), each computed at two different scales (1 and 2). The only difference is that scale 2 algorithms sub-sample the image by two. We will analyze both the overall metric and all six parameters: *GM1*, *RO1*, *RM1*, *GM2*, *RO2* and *RM2*.

Video Intrinsic Integrity and Distortion Evaluation Oracle (VIIDEO) [19] is a VQA that analyzes statistical regularities in natural videos, and yields zero for images and still videos.

This report omits several IQA and VQA metrics mentioned in literature. The University of Texas at Austin has published a series of metrics that build upon each other and improve from one generation to the next, and the newer metrics are of more interest. The VQEG Image Quality Evaluation Tool (VIQET) was omitted because the user interface was incompatible with our batch processing software, an issue we did not have time to address. FRIQUEE [20] was omitted due to its unduly slow speed.

Table 1. NR Metrics

Metric	**Description**	**Ref.**	**Type**
BRISQUE	Blind / referenceless image spatial quality evaluator	[2]	IQA
CurveletQA	Curvelet quality assessment	[16]	IQA
IL-NIQE	Integrated local NIQE	[17]	IQA
NIQE	Naturalness image quality evaluator	[1]	IQA
OG-IQA	Oriented gradients image quality assessment	[18]	IQA
QAC	Quality-aware clustering	[14]	IQA
SSEQ	Spatial-spectral entropy-based quality	[15]	IQA
VIIDEO	Video intrinsic integrity and distortion evaluation oracle	[19]	VQA

5. METRIC ANALYSIS ON UNSEEN DATA

Each of the NR metrics in Section 4 was designed for a niche application. The challenge we address in this paper is moving from silos to a broadly applicable metric. This change of scope will decrease metric performance. Machine learning metrics were not retrained, because this is beyond the capabilities of a typical user. Moreover, retraining machine learning metrics would create a chicken and egg situation. By retraining the metric on our datasets, we would invalidate our performance analyses.

Let us start by comparing metric performance on known data and unseen data. Figure 1 and Table 2 compare the reported performance of each NR metric and NR parameter (from the original publication) to our unseen camera data (*cam90%*) and our unseen video codec data (*ITS4S* and *vqegHDcuts*). A range of values appears when the original publication included two or more analyses. IQAs were applied to the video sequences on a frame-by-frame basis and the mean taken over time. Interlaced videos were de-interlaced by field duplication. Images were converted to 4 s still videos for VIIDEO, but VIIDEO yielded zero; the impacted VIIDEO analyses are omitted. The performance of IL-NIQE and SSEQ on *vqegHDcuts* is unavailable, due to slow run speed.

Figure 1 and Table 2 show a large drop in metric accuracy as we move from the reported performance to unseen data. Select scatter plots are provided for deeper understanding. Figures 2 and 3 show scatter plots between NR metrics the unseen camera data (*cam90%*) and the unseen video codec data (*ITS4S*) respectively. When interpreting this data, recall that *cam90%* and *vqegHDcuts* use techniques that are unprecedented, untested, and a source of error. This data is provided to show trends, as we lack better validation data.

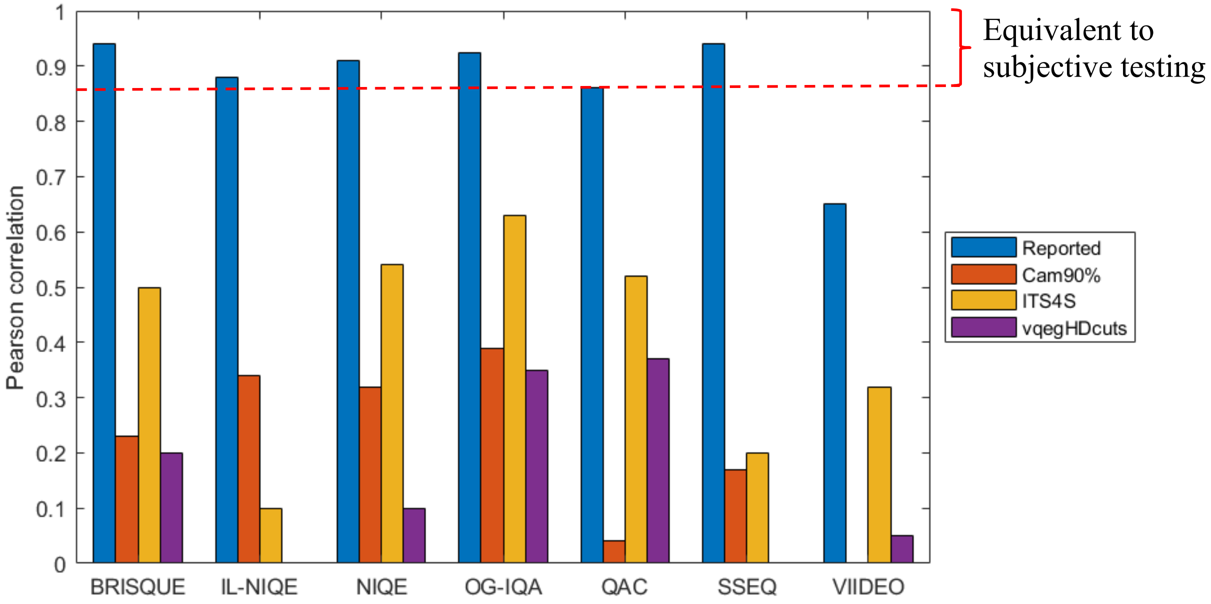

Figure 1. Comparison between reported metric performance and unseen data, using Pearson correlation, presented as a histogram.

Table 2. Comparison between reported metric performance and unseen data, using Pearson correlation.

NR Metric	***Reported Performance***	***Unseen Data***		
		cam90%	*ITS4S*	*vqegHDcuts*
BRISQUE	0.94	-0.23	**-0.50**	-0.20
CurveletQA	0.93	—	—	—
IL-NIQE	0.88	**-0.34**	-0.10	—[2]
NIQE	0.91	**-0.32**	**-0.54**	-0.10
OG-IQA	0.90..0.95	**-0.39**	**-0.63**	**-0.35**
QAC	0.84..0.88	0.04	**0.52**	**0.37**
SSEQ	0.94	-0.17	-0.20	—[2]
VIIDEO	0.65	—[3]	**-0.32**	-0.05
NR Parameters from CurveletQA				
f1	—	-0.23	-0.07	-0.16
f2	—	**0.39**	**0.36**	0.25
f3	—	**0.32**	0.13	-0.19
f4	—	**0.42**	**0.32**	0.14
f5	—	-0.17	-0.09	**-0.38**
f6	—	-0.06	-0.12	-0.26
f7	—	-0.06	**-0.47**	0.14
f8	—	-0.12	**-0.49**	-0.29
f9	—	-0.09	0.07	-0.28
f10	—	-0.14	0.07	0.00
f11	—	-0.11	**-0.56**	-0.05
f12	—	-0.12	-0.26	**-0.32**
NR Parameters from OG-IQA				
GM1	0.74	**-0.42**	-0.04	-0.09
RO1	0.92	-0.23	**-0.61**	**-0.32**
RM1	0.76	**-0.49**	-0.25	-0.04
GM2	0.74	**-0.36**	0.03	-0.09
RO2	0.92	-0.15	-0.28	**-0.34**
RM2	0.76	**-0.42**	-0.05	-0.17

[2] Unavailable due to slow metric code run speed.
[3] VIIDEO yields 0 for images, which biases the *cam90%* analysis.

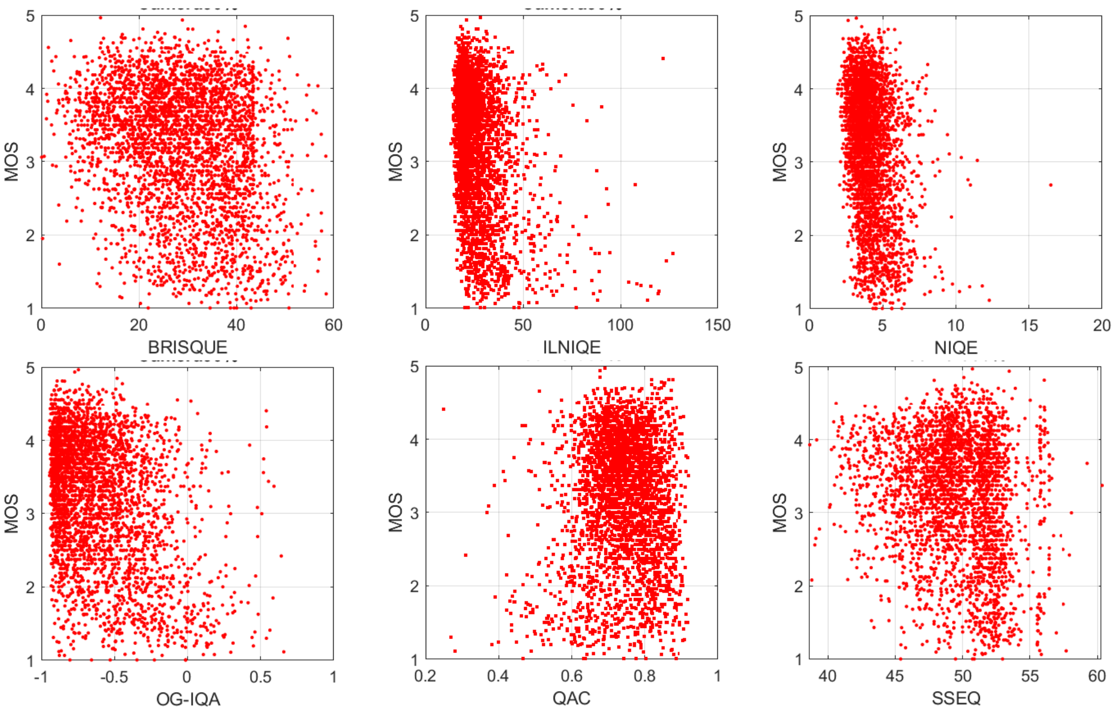

Figure 2. Performance of IQA NR metrics on the *cam90%* data.

The statistics in Table 2 are color coded to simplify interpretation. **Bold blue** indicates 0.32 or higher Pearson correlation magnitude between the metric and unseen data. These metrics explain at least 10% of the variance in MOSs (given by correlation squared). Red indicates 0.1 or lower correlation magnitude between the metric and unseen data, aka less than 1% of the variance in MOSs. In the “reported” column, purple indicates a reported performance of 0.85 or higher, which is equivalent to subjective testing. Such claims require extraordinary proof and may indicate over-training.

The 0.85 threshold is based on VQEG’s lab-to-lab comparisons [21] and VQEG’s analysis of the accuracy of subjective experiments with five subjects [22]. Section 8 of [21] gives robust estimates based on common video sequences inserted into 41 experiments. The Pearson correlations ranged from 0.94 to 0.996. This estimate assumes a carefully balanced set of SRC, a wide range of HRC quality and 24 subjects. VQEG’s study on the influence of subject and environment on subjective testing indicates trends for less carefully designed experiments [22]. With 5 subjects, Pearson correlations ranged from 0.85 to 0.96, plus a long tail of lower correlation outliers. Based on these statistics, we chose a threshold of 0.85.

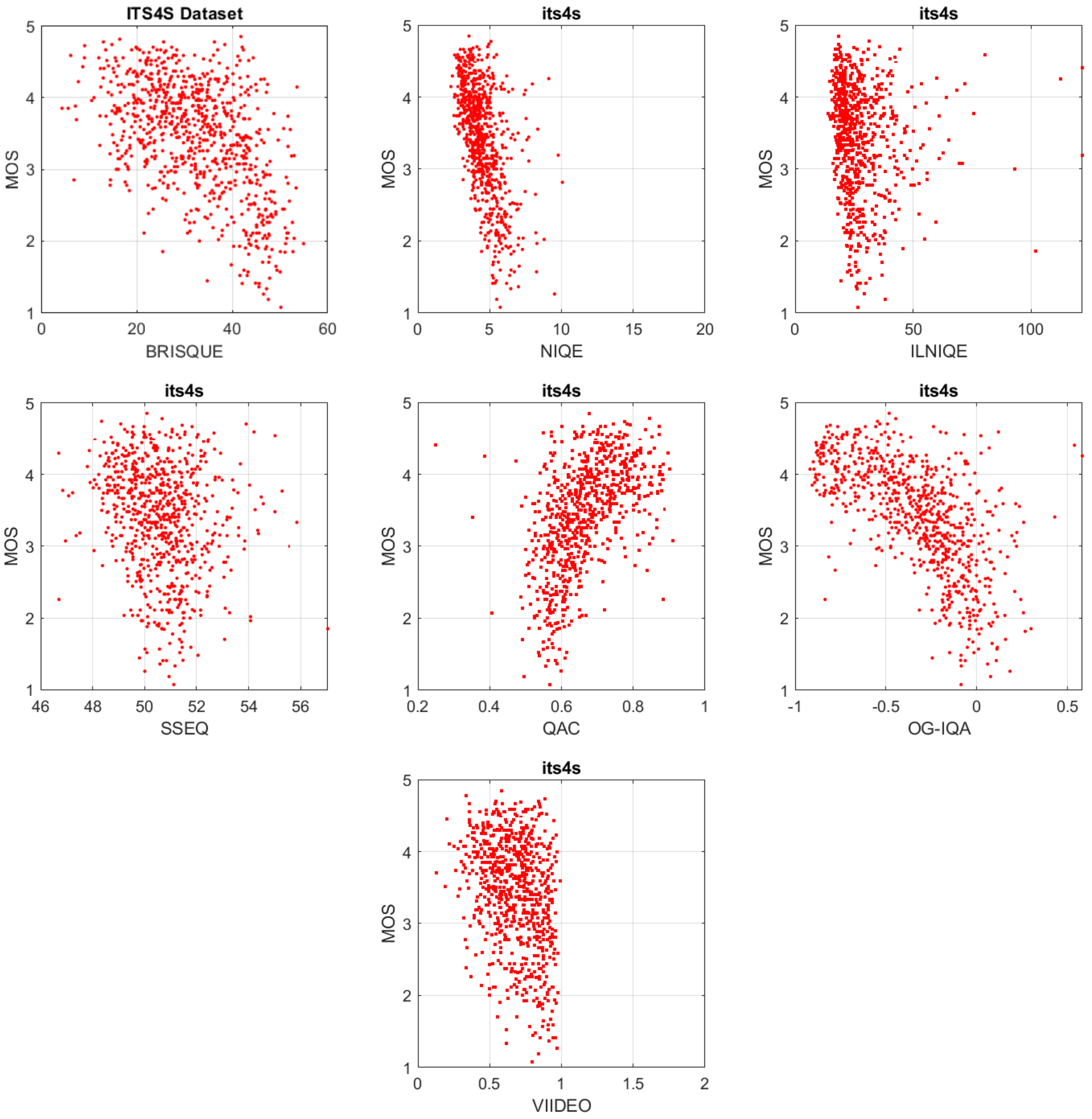

Figure 3. Response of IQA and VQA NR metrics to falling bitrate.

Table 3 splits *cam90%* into individual datasets, for a more precise evaluation of camera capture impairments. To perform well, the NR metrics must predict the quality impact of the camera capture, the camera operator, and scene aesthetics. VIIDEO computes temporal differences, and so can only be compared to the *ITS4S* dataset's SRC.

Table 4 shows Pearson correlation between each NR metric and *ITS4S*, split by HRC. Notice how each metric's performance changes as we move from camera impairments (*ITS4S* SRC) to compression artifacts (H.264 at 512 Kbps). Each HRC spans a limited range of MOSs, so the Pearson correlation values are reduced (see [22] for details). Table 4 shows complex responses that may require additional investigation. For example, *NIQE* has relatively stable correlations, while *VIIDEO* and *OG-IQA RM1* changes in magnitude and sign.

The statistics in these three tables fail to show an obvious performance advantage of the NR metrics over the NR parameters. The NR metrics that use machine learning would ideally be re-trained, but that is impractical for most people. The NR parameters were always intended to operate within a larger metric. This seems a more viable strategy, which allows disparate researchers to solve different portions of the NR metric problem. A robust NR metric would contain a variety of NR parameters that focus on different aspects of quality. Still, this approach would pose significant challenges for machine learning (e.g., how to create large datasets of suitable training data).

We were uncomfortable replicating MOSs to create *vqegHDcuts*. To evaluate the impact of this procedure, we will create two subsets. For each sequence in the unmodified VQEG datasets, subset *Vmax* retains the segment that maximizes the NR metric rating, and subset *Vmin* retains the segment that minimizes the NR metric rating. We then calculate Pearson correlation between the corresponding NR ratings for *Vmin* and *Vmax*. These correlations are 0.77 for NIQE, 0.98 for OG-IQA and 0.79 for QAC. (NIQE, OG IQA, and QAC are the top three NR metrics in Table 2, based on performance on the *ITS4S* dataset.) When we calculate Pearson correlation between MOSs and either the *Vmin* or *Vmax* subset, all values are within +/- 0.01 of the values shown in Table 2. We conclude that *vqegHDcuts*'s replicated MOSs are close enough to the true MOSs.

This technique must be used with caution. We used the cutting technique to create a faux dataset from [23], and the results were not useful. The likely problem is that the dataset uses 15 s SRC with rapid scene cuts and obvious temporal changes in motion and coding complexity. The differences between *Vmax* and *Vmin* were large (e.g., 10% of the NR metric's range on average), and the apparent accuracy of all NR metrics plummeted. The problem appeared to be associated with the dataset, not the metrics.

Table 3. NR metric performance for camera capture.

NR Metric	***ITS4S SRC***	***ITS4S2***	***CCRIQ***	***C&V***	***Wild***
BRISQUE	-0.20	-0.20	-0.28	-0.29	-0.21
IL-NIQE	-0.12	-0.28	**-0.38**	**-0.34**	**-0.49**
NIQE	-0.17	**-0.35**	**-0.37**	**-0.42**	-0.19
OG-IQA	-0.13	**-0.35**	**-0.54**	**-0.58**	**-0.33**
QAC	-0.12	-0.14	**0.32**	0.03	0.06
SSEQ	0.21	-0.14	-0.02	-0.13	-0.26
VIIDEO	0.08	—	—	—	—
NR Parameters from CurveletQA					
f1	-0.01	-0.16	-0.30	-0.26	-0.23
f2	0.12	0.31	**0.53**	**0.45**	**0.39**
f3	0.18	**0.48**	**0.40**	0.14	**0.32**
f4	0.11	**0.39**	**0.55**	**0.48**	**0.42**
f5	-0.03	-0.12	-0.19	-0.07	-0.17
f6	-0.02	0.02	-0.06	-0.08	-0.06
f7	-0.03	0.01	-0.20	**-0.34**	-0.06
f8	0.00	-0.03	-0.26	-0.28	-0.12
f9	0.09	-0.01	-0.15	-0.17	-0.09
f10	-0.12	-0.16	-0.02	0.02	-0.14
f11	-0.02	-0.01	-0.27	**-0.37**	-0.11
f12	0.05	-0.03	-0.23	-0.25	-0.12
NR Parameters from OG-IQA					
GM1	-0.16	**-0.42**	**-0.47**	**-0.36**	**-0.43**
RO1	0.04	-0.09	**-0.44**	**-0.45**	-0.26
RM1	-0.20	**-0.49**	**-0.53**	**-0.43**	**-0.49**
GM2	-0.15	**-0.35**	**-0.37**	**-0.34**	**-0.37**
RO2	0.12	0.00	**-0.34**	**-0.30**	-0.27
RM2	-0.11	**-0.41**	**-0.49**	**-0.36**	**-0.43**

Table 4. NR metric response to dropping bit-rate

NR Metric	***ITS4S SRC***	***2.34 Mbps***	***1.732 Mbps***	***1.256 Mbps***	***951 Kbps***	***512 Kbps***
BRISQUE	-0.20	-0.15	-0.21	-0.23	**-0.35**	**-0.45**
IL-NIQE	-0.12	0.02	0.04	-0.09	0.08	0.10
NIQE	-0.17	-0.14	-0.28	**-0.38**	**-0.37**	-0.27
OG-IQA	-0.13	-0.10	-0.14	-0.26	-0.22	-0.24
QAC	-0.12	-0.04	0.09	0.11	0.08	0.06
SSEQ	0.21	0.03	0.14	0.08	0.21	0.15
VIIDEO	0.08	0.22	0.08	-0.18	**-0.34**	**-0.34**
NR Features from CurveletQA						
f1	-0.01	0.10	-0.06	0.26	0.15	0.18
f2	0.12	0.09	-0.09	**-0.33**	**-0.45**	**-0.46**
f3	0.18	0.15	-0.06	-0.21	**-0.42**	**-0.42**
f4	0.11	0.07	-0.08	**-0.33**	**-0.42**	**-0.47**
f5	-0.03	-0.09	0.18	0.06	0.11	0.16
f6	-0.02	-0.18	0.19	-0.06	-0.19	-0.15
f7	-0.03	0.04	-0.28	-0.14	-0.30	**-0.51**
f8	0.00	-0.17	-0.02	0.09	-0.09	-0.24
f9	0.09	-0.05	0.17	0.26	**0.36**	0.31
f10	-0.12	-0.01	0.12	0.07	0.23	0.28
f11	-0.02	-0.08	-0.23	-0.09	-0.27	**-0.46**
f12	0.05	-0.11	0.10	0.22	0.19	0.04
NR Features from OG-IQA						
GM1	-0.16	-0.09	0.04	0.01	0.24	**0.32**
RO1	0.04	0.03	-0.03	0.03	-0.23	**-0.37**
RM1	-0.20	-0.13	0.04	-0.01	0.30	**0.33**
GM2	-0.15	-0.06	0.05	0.02	0.20	**0.33**
RO2	0.12	0.04	0.11	0.17	0.16	0.11
RM2	-0.11	-0.08	0.05	0.02	0.28	**0.36**

Few of the publicly available IQA and VQA subjective test datasets focus on consumer camera applications. Due to this lack of unforeseen data, this paper uses unprecedented techniques to combine datasets (i.e., *cam90%* and *vqegHDcuts*). Analyses of these combined datasets provide insights into overall performance. The magnitude of error cannot be measured, but generally we must assume the metrics perform better than reported. The individual dataset analyses do not have this problem, but the opposite caution applies. *ITS4S*, *ITS4S2*, *CCRIQ*, *C&V*, and *Wild* each have a limited scope, so we must assume that chance plays a large role. When interpreting the analyses in this section, remember that our goal is not independent metric validation in support of an international standard. These analyses were designed to a lower standard, that of deciding whether existing NR-IQA and NR-VQA metrics meet industrial and consumer needs around consumer camera applications.

6. CONCLUSION

Our analysis shows the need for more research and development on NR metrics for IQA and VQA. The metrics we examined developed problems when moved from a narrow use case into the broad application of consumer content. All of these NR metrics had reduced accuracy when given previously unseen stimuli. NR metrics trained on traditional distortions (e.g., white noise, Gaussian blur) and traditional experiment designs (e.g., limited source scenes) did not generalize to unseen distortions. Camera impairments are particularly problematic.

We present a set of six datasets, all openly available, that are appropriate for use in developing IQA and VQA metrics. These datasets include samples of entertainment content and public safety content, created by a variety of professional and amateur camera operators. The datasets emphasize diverse subject matter from a variety of consumer cameras, depicting problems associated with the camera capture as well as problems associated with compression.

7. REFERENCES

[1] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a Completely Blind Image Quality Analyzer ," *IEEE Signal processing Letters*, pp. 209-212, vol. 22, no. 3, March 2013.

[2] A. Mittal, A. K. Moorthy and A. C. Bovik, "Referenceless Image Spatial Quality Evaluation Engine," *45th Asilomar Conference on Signals*, Systems and Computers. November 2011.

[3] H. R. Sheikh, M. F. Sabir, A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Quality Assessment Algorithms", *IEEE Transactions on Image Processing* , vol. 15, no. 11, pp. 3440-3451, November 2006.

[4] L. Janowski, L. Malfait, and M. Pinson, "Evaluating experiment design with unrepeated scenes for video quality subjective assessment," *Springer Quality of User Experience*, June 2019.

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

[6] Laboratory for Image & Video Engineering (LIVE), University of Texas at Austin, <http://live.ece.utexas.edu/research/quality/index.htm>, accessed on January 30, 2019.

[7] Consumer Digital Video Library (CDVL), <https://www.cdvl.org/>, accessed on Jan. 30, 2019.

[8] Margaret H. Pinson, "ITS4S2: an image quality dataset with unrepeated images from consumer cameras," NTIA Technical Memo TM-19-537, April 2019.

[9] "VIME Image Database," <https://www.flickr.com/groups/vime/>.

[10] 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 August 2015, pp. 1-6.

[11] M. A. Saad et al., "Image Quality of Experience: A Subjective Test Targeting the Consumer's Experience," *Proceedings of the International Symposium on Electronic Imaging 2016, Human Vision and Electronic Imaging 2016*, February 14, 2016.

[12] J. Nawała et al., "Image quality datasets for consumer camera evaluation," publication pending.

[13] D. Ghadiyaram and A.C. Bovik, "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality," *IEEE Transactions on Image Processing*, Journal of Vision, vol. 17, no. 1, pp. 1-25, 2017.

[14] W. Xue, L. Zhang and X. Mou, "Learning without human scores for blind image quality assessment," *IEEE Conference on Computer Vision and Pattern Recognition*, June 2013.

[15] L. Liu, B. Liu, H. Huang, and A. C. Bovik "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, June 2014.

[16] Lixiong Liu, Hongping Dong, Hua Huang, Alan C. Bovik, " No-reference image quality assessment in curvelet domain," *Signal Processing: Image Communication,* February 2014.

[17] L. Zhang, L. Zhang and A.C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, August 2015.

[18] L. Liu, Y. Hua, Q. Zhao, H. Huang and A.C. Bovik, "Blind image quality assessment by relative gradient statistics and Adaboosting neural network," *Signal Processing: Image Communication*, vol. 40, no. 1, pp. 1-15, January, 2016.

[19] A. Mittal, M. A. Saad, and A. C. Bovik, "A Completely Blind Video Integrity Oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289-300, January, 2016.

[20] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 1-25, 2017.

[21] A. Webster et al., "Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase I," VQEG, March 28, 2008.

[22] M. Pinson et al., "The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 6, No. 6, October 2012.

[23] Margaret H. Pinson; Stephen Wolf; Gregory W. Cermak, "HDTV Subjective Quality of H.264 vs. MPEG-2, with and without Packet Loss," *IEEE Transactions on Broadcasting*, vol.56, no.1, pp.86-91, March 2010.

NTIA FORM 29
(4-80)

U.S. DEPARTMENT OF COMMERCE
NATIONAL TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO. TM-20-547	2. Government Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE
Analysis of No-Reference Metrics for Image and Video Quality of Consumer Applications

5. Publication Date
January 23, 2019

6. Performing Organization Code

7. AUTHOR(S)
Margaret H. Pinson

9. Project/Task/Work Unit No.

6784000-300 and 6860000-303

8. PERFORMING ORGANIZATION NAME AND ADDRESS
Institute for Telecommunication Sciences
National Telecommunications & Information Administration
U.S. Department of Commerce
325 Broadway
Boulder, CO 80305

10. Contract/Grant Number.

11. Sponsoring Organization Name and Address
Public Safety Communications Research (PSCR) Division
Communications Technology Laboratory (CTL)
National Institute of Standards and Technology (NIST)
U.S. Department of Commerce
325 Broadway
Boulder, CO 80305

12. Type of Report and Period Covered

14. SUPPLEMENTARY NOTES

15. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)

This paper analyzes the performance of eight no-reference (NR) metrics. Seven assess image quality (BRISQUE, CurveletQA, IL-NIQE, NIQE, OG-IQA, QAC and SSEQ) and one assesses video quality (VIIDEO). The challenge we address in this paper is moving from research silos to a broadly applicable metric. Our analyses use six new subjective datasets that characterize modern cameras and high performing networks. Five datasets were designed around consumer applications and no-reference metric development. The sixth dataset was designed for full-reference metric analyses; we present a technique to modify older datasets for NR metric development. Our analyses show a need for more research and development. The NR metrics were inaccurate for consumer applications.

16. Key Words (Alphabetical order, separated by semicolons)

BRISQUE, CurveletQA, IL-NIQE, image quality, NIQE, no-reference metrics, NR, OG-IQA, QAC, SSEQ, video quality, VIIDEO

17. AVAILABILITY STATEMENT ☒ UNLIMITED. ☐ FOR OFFICIAL DISTRIBUTION.	18. Security Class. (This report) Unclassified	20. Number of pages 28
	19. Security Class. (This page) Unclassified	21. Price: N/A

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

This report receives both local NTIA and other agency review. Both agencies' logos and report series numbering appear on the cover.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail ITSinfo@ntia.gov.