CROWDSOURCING SUBJECTIVE IMAGE QUALITY EVALUATION

Flávio Ribeiro[1]*, Dinei Florencio*[2] *and Vítor Nascimento*[1]

[1] Electronic Systems Engineering Department, Universidade de São Paulo, Brazil
[2] Microsoft Research, One Microsoft Way, Redmond, WA, 98052

ABSTRACT

Subjective tests are generally regarded as the most reliable and definitive methods for assessing image quality. Nevertheless, laboratory studies are time consuming and expensive. Thus, researchers often choose to run informal studies or use objective quality measures, producing results which may not correlate well with human perception. In this paper we propose a cost-effective and convenient subjective quality measure called crowdMOS, obtained by having internet workers participate in MOS (mean opinion score) subjective quality studies. Since these workers cannot be supervised, we propose methods for detecting and discarding inaccurate or malicious scores. To facilitate this process, we offer an open source set of tools for Amazon Mechanical Turk, which is an internet marketplace for crowdsourcing. These tools completely automate the test design, score retrieval and statistical analysis, abstracting away the technical details of Mechanical Turk and ensuring a user-friendly, affordable and consistent test methodology. We demonstrate crowdMOS using data from the LIVE subjective quality image dataset, showing that it delivers accurate and repeatable results.

Index Terms— crowdsourcing, subjective quality, quality assessment, mean opinion score, MOS, mechanical turk.

1. INTRODUCTION

Quality measures are of crucial importance for benchmarking, tuning and monitoring signal processing algorithms. Since multimedia systems typically have human end-users, the definitive quality measure is given by human perception. Despite many efforts to design accurate objective measures, so far none have been able to account for all the peculiarities of human physiological and psychological responses. Thus, subjective tests are generally regarded as the most reliable methods for assessing image quality.

One of the most popular subjective measures is the absolute category rating (ACR), which was standardized for images and video in ITU-T P.910 [1]. It consists of having a panel of volunteers rate samples under controlled conditions, using the discrete 1-5 scale shown in Table 1. Since the score for a particular sample is given by its average over all subjects, this is also known as the MOS (mean opinion score) test. Variations of the MOS test have been developed for several specific domains [2, 3].

In general, standardized subjective tests require:

1. recruiting a panel of 20-40 volunteers of sufficient diversity to deliver statistically significant results;
2. conducting experiments in a laboratory environment, equipped with hardware conforming to the test standard;
3. providing each volunteer with the same sets of stimuli and instructions.

The degree to which these requirements are followed strongly determines the accuracy and repeatability of a MOS test. Unfortunately, laboratory testing is too expensive and time consuming for frequent use. Thus, researchers often perform informal MOS tests with smaller panels, or with individuals not drawn from the general population, or use objective quality measures (e.g. [4, 5, 6]).

Table 1. MOS (ACR) scores

Rating	Quality	Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

Most objective quality measures for images are full-reference, since they require a perfect quality original as a basis for comparison. Nevertheless, quality often depends on characteristics which are difficult to infer, even given an undistorted reference. For instance, visual attention and gaze direction are known to significantly influence subjective quality [7].

Objective quality measures are convenient because they do not have the costs associated with human subjects. However, each objective measure is designed to estimate specific quality aspects and is tuned to map to a unique dataset. Thus, an objective measure will only produce predictable results for the environment, error conditions and impairments it was developed for. This sensitivity is especially severe for no-reference measures, which are typically developed to detect one specific impairment.

In this paper we propose crowdMOS, a family of subjective quality measures obtained by relaxing requirement (2) listed above. Instead of running a MOS test in a controlled environment, we outsource this task to workers from an internet crowd, who evaluate the quality of multimedia files at their leisure, using their own hardware. The recruiting process is performed using Amazon Mechanical Turk (MTurk), an internet marketplace where hundreds of thousands of workers get paid to complete small tasks which require relatively little training. To automate the experiment design and hide technical details involving MTurk, we developed a set of open-source, cross-platform tools[1] [8].

Since MTurk workers are unsupervised, the determination of accurate confidence intervals (CIs) is particularly important. While performing a literature review, we noticed that it is not widely acknowledged how to compute CIs for algorithm scores, which are affected by preference variations across users, intrinsic quality variations across test files and subjective uncertainty. In fact, many publications present no CIs, or make incorrect assumptions about the independence of scores, producing unreasonably optimistic CIs. Thus, a useful byproduct of crowdMOS is a two-way random effects model to model uncertainty in subjective tests, allowing the determination of accurate CIs and the planning of experiments with adequate sample sizes.

Even though crowdsourcing has become a popular approach for conducting surveys and user studies, to our knowledge [9] is the only previous work related to subjective quality assessment (QA). In con-

[1] Available online at http://research.microsoft.com/crowdmos/.

Fig. 1. Interface for image quality assessment

trast to crowdMOS, this approach is based on binary paired comparison. Our preference for MOS is threefold: (i) MOS and its differential variants are de facto standards for subjective QA, and we prefer to maintain compatibility with previous work; (ii) by using a 5 point scale, MOS can obtain more accurate results with a smaller number of scores; (iii) with MOS, workers can be instructed on the meaning of scores, allowing for more diverse designs. Finally, crowdMOS is unique since it offers an open-source set of tools, which allows any researcher to perform their own tests without requiring our intervention.

As we show in the following sections, crowdMOS can be used to produce accurate and repeatable scores with costs which are orders of magnitude lower than those for laboratory testing. Given the large pool of MTurk workers, panel sizes can be scaled arbitrarily with no additional overhead. MTurk workers also represent a less biased sample of the general population than the student volunteers who are often recruited for academic research. By combining the proposed statistical analysis with the scalability of crowdsourcing, one can design studies to an arbitrary level of accuracy, which would be impractical using traditional means.

Due to limited space, our examples and language are focused on image QA. In previous work, we have detailed the use of crowdMOS for speech QA [10], and our methodology can be easily extended for testing many other types of multimedia content.

2. CROWDSOURCING AND MECHANICAL TURK

Amazon MTurk is an online marketplace where workers can complete human intelligence tasks (HITs) in exchange for monetary rewards. In addition to the base pay associated with a HIT, a worker can also receive a bonus for outstanding work.

MTurk tasks should be designed to be as simple as possible. A typical MTurk HIT requires no more than a few minutes to complete, and requires no specialized training. Since MTurk accepts workers from all over the world, the typical pay is below minimum wage in the United States.

For image quality HITs, we designed an interface using dynamic HTML, which is shown on Fig. 1. It displays one image at a time, which saves screen real-estate, and is useful for detecting faint variations in multi-stimulus tests, where one compares a clean reference with one or more degraded versions. Our HITs require between one and two minutes to complete, and feature around 10 images.

By using the crowdMOS tools, the researcher does not have to interact directly with MTurk. The tools come with templates suitable for implementing the various standardized tests described in section 3.1. Thus, the researcher only has to collect a list of files to evaluate and set basic HIT parameters (such as a title, description and reward). The tools can then be used to automatically create HITs and later retrieve, update, analyze and approve submitted assignments.

Since MTurk is a marketplace, requesters compete among themselves for the attention of the workers. To ensure a high throughput and good quality results, a requester has to provide appropriate incentives, low barriers for entry and design engaging experiments. For instance, a requester can require workers to pass qualification tests before working on their HITs. Our experience is that for subjective QA, using qualifications only shrinks the worker pool without a measurable increase in quality. Indeed, qualification tests are not rewarded, such that our relatively short tasks do not provide sufficient incentive for most users to take them. Since our qualification was limited to instructions and very easy test, it did little to improve quality.

Our studies offered a base reward of $.05/HIT. To keep workers from quitting early and to promote high quality results, we paid a bonus of $.05/HIT if a user submitted the full set of HITs comprising a study. Among the users who submitted full sets, we ranked their results according to correlation to the mean, and paid an additional $.05/HIT if the user was in the top 50%, or an additional $.10/HIT if the user was on the top 10%. Assuming a mean working time of 2 minutes/HIT, the average pay is around $5.70/hour. Given the international pool of workers, this is sufficient to run large studies with hundreds of images in only a few hours.

3. EXPERIMENT DESIGN AND ANALYSIS

3.1. Designs for image quality assessment

A given subjective quality study is meant to detect and measure specific distortions which must be within the discrimination capacity of the subjects. Most standards for multimedia QA were first drafted in the early 1990s or earlier, when professional-grade equipment was imperative to ensure consistent high quality reproduction. Current consumer-level products are able to reproduce multimedia content with accuracy which well exceeds the requirements for intermediate quality applications, enabling a wide variety of studies.

CrowdMOS implements the following test methodologies:

- A basic single-stimulus ACR test [1], where each HIT is created by drawing a fixed number of files from the sample pool, without replacement.
- A single-stimulus ACR test with the constraint that a HIT never contains two samples created from the same test signal. This constraint promotes ACR as an absolute scale by not tempting users to make relative judgements.
- A basic multiple-stimulus ACR test, where each HIT only has samples created from the same test signal. This allows workers to discriminate finer differences.
- The MUSHRA (multiple-stimulus with hidden reference and anchor) test [11], where the reference is presented to the subject before the test, and the same reference and an anchor are hidden among the test files. The anchor is produced by processing the reference by an algorithm of known degradation, and provides a baseline to minimize distortions of the score chart due to the use of relative comparisons.
- The DSIS (double-stimulus impairment scale) test [3], where each HIT only presents a reference and its processed version. Both images are labeled as such, and the subject must evaluate how much degradation was caused by the processing.

3.2. Confidence intervals and worker screening

Since MTurk workers are unsupervised, postscreening is necessary to identify and remove outliers which could affect the accuracy of the study. Fortunately, workers have little incentive to submit malicious results, because their approval percentages are used as a qualification requirement by many requesters. Providing bonuses and paying adequate rewards also helps to promote quality. Nevertheless, large studies invariably have at least one worker who submits seemingly random results.

In laboratory studies, volunteers can be instructed to always score a precise subset of samples. In crowdMOS, we have no such choice, since HITs are assigned randomly and a worker can quit at any time. Thus, the following proposal is designed to ignore missing values.

We first address the algorithm comparison problem of evaluating K algorithms using M reference images. All references are processed by all algorithms, producing KM samples which are rated by N workers. In this case, we are interested in the MOS and CI for each algorithm. We then consider an image comparison problem, where there is no concept of algorithm and we simply want to obtain the MOS and CI for M images using N workers.

3.2.1. Determining confidence intervals

Consider a fixed algorithm of interest whose mean score μ we wish to estimate. Let $\mu_{m,n}$ be the score given to image m by worker n, with $1 \leq m \leq M$ and $1 \leq n \leq N$. To determine CIs for a wide variety of experiments, we use the two-way random effects model given by

$$\mu_{m,n} = \mu + \alpha_m + \beta_n + \varepsilon_{m,n}$$
$$\alpha_m \sim \mathcal{N}\left(0, \sigma_s^2\right), \quad \beta_n \sim \mathcal{N}\left(0, \sigma_w^2\right), \quad \varepsilon_{m,n} \sim \mathcal{N}\left(0, \sigma_u^2\right),$$

where σ_s^2, σ_w^2, and σ_u^2 model the diversity of intrinsic sample quality, diversity of worker preference, and subjective uncertainty (σ_s^2, σ_w^2 and σ_u^2 depend on the algorithm). To simplify this discussion and the notation, assume there are no missing scores. We then have the estimates

$$\begin{aligned}
\hat{\mu} &= \tfrac{1}{MN}\textstyle\sum_{n=1}^{N}\sum_{m=1}^{M}\mu_{m,n} \\
\hat{\sigma}_w^2 + \hat{\sigma}_u^2 &= \tfrac{1}{M}\textstyle\sum_{m=1}^{M}\operatorname{var}\left(\mu_{m,1}, ..., \mu_{m,N}\right) \\
\hat{\sigma}_s^2 + \hat{\sigma}_u^2 &= \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\operatorname{var}\left(\mu_{1,n}, ..., \mu_{M,n}\right) \\
\hat{\sigma}_s^2 + \hat{\sigma}_w^2 + \hat{\sigma}_u^2 &= \operatorname{var}\left(\mu_{1,1}, ..., \mu_{M,N}\right).
\end{aligned}$$

$\hat{\sigma}_s^2$, $\hat{\sigma}_w^2$ and $\hat{\sigma}_u^2$ can be obtained from the above using a least-squares estimate (if there are missing scores, $\hat{\sigma}_w^2 + \hat{\sigma}_u^2$ and $\hat{\sigma}_s^2 + \hat{\sigma}_u^2$ can be determined by averaging sample variances over smaller blocks of fixed size). An estimate of the mean score variance is given by

$$\operatorname{var}\left[\hat{\mu}\right] = \tfrac{\hat{\sigma}_s^2}{M} + \tfrac{\hat{\sigma}_w^2}{N} + \tfrac{\hat{\sigma}_u^2}{MN}.$$

To exactly determine the 95% CI for $\hat{\mu}$, one must integrate the PDF of the sum of 3 scaled t-distributed random variables with $M-1$, $N-1$ and $MN-1$ degrees of freedom, which is quite inconvenient to determine. Instead, the crowdMOS tools use a slightly more conservative CI for $\hat{\mu}$ given by

$$\left[\hat{\mu} - t\sqrt{\operatorname{var}\left[\hat{\mu}\right]}, \hat{\mu} + t\sqrt{\operatorname{var}\left[\hat{\mu}\right]}\right],$$

where t is the appropriate percentile from a t distribution with $\min\left(N, M\right) - 1$ degrees of freedom.

To validate this approach, we used an experiment where all the scores were available and compared the obtained CIs with those produced by percentile bootstrap resampling [12] (a non-parametric method), with very similar results. Our approach is much more convenient, since unlike bootstrap resampling, it can be easily extended to work with missing values and does not require a computationally intensive procedure.

The problem of determining the MOS and CI for each image can be treated by defining $\mu_m = \mu + \alpha_m$ as the true MOS for image m, and making β_n and $\varepsilon_{m,n}$ image dependent. For each m, we have the estimates

$$\hat{\mu}_m = \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\mu_{m,n}, \quad \hat{\sigma}_{w,m}^2 + \hat{\sigma}_{u,m}^2 = \operatorname{var}\left(\mu_{m,1}, ..., \mu_{m,N}\right).$$

It can be shown that since each $\mu_{m,n}$ is normally distributed, the CI for $\hat{\mu}_m$ is given by

$$\left[\hat{\mu}_m - t\sqrt{\left(\hat{\sigma}_{w,m}^2 + \hat{\sigma}_{u,m}^2\right)/N}, \hat{\mu}_m + t\sqrt{\left(\hat{\sigma}_{w,m}^2 + \hat{\sigma}_{u,m}^2\right)/N}\right],$$

where t is the appropriate percentile from a t distribution with $N-1$ degrees of freedom.

3.2.2. Post-screening

Once enough scores have been submitted, we screen workers. For the algorithm comparison problem, we compute the global MOS values $\hat{\mu}^k$ and the worker MOS values $\hat{\nu}_n^k = \frac{1}{M}\sum_{m=1}^{M}\mu_{m,n}^k$, for $1 \leq k \leq K$ and $1 \leq n \leq N$. Let

$$r_n = \frac{\operatorname{cov}\left(\hat{\mu}^1, ..., \hat{\mu}^K; \hat{\nu}_n^1, ..., \hat{\nu}_n^K\right)}{\sqrt{\operatorname{var}\left(\hat{\mu}^1, ..., \hat{\mu}^K\right)}\sqrt{\operatorname{var}\left(\hat{\nu}_n^1, ..., \hat{\nu}_n^K\right)}},$$

which is the Pearson sample correlation coefficient between the MOS estimates from worker n and the global MOS estimates. If $r_n < 0.25$ (which is an arbitrarily chosen, conservative threshold), all HITs from worker n are rejected. All r_n values are recomputed for the remaining HITs, and workers are ranked in decreasing order of r_n. Workers are then awarded the bonuses described in Section 2.

For the image comparison problem, we define instead

$$r_n = \frac{\operatorname{cov}\left(\hat{\mu}_1, ..., \hat{\mu}_M; \mu_{1,n}, ..., \mu_{M,n}\right)}{\sqrt{\operatorname{var}\left(\hat{\mu}_1, ..., \hat{\mu}_M\right)}\sqrt{\operatorname{var}\left(\mu_{1,n}, ..., \mu_{M,n}\right)}},$$

and perform thresholding as shown in the previous paragraph.

To compute final MOS and CI values for both the algorithm and image comparison problems, a second level of screening is used. We first normalize worker scores such that

$$z_{m,n} = \frac{\mu_{m,n} - \hat{\mu}_m}{\sqrt{\hat{\sigma}_{w,m}^2 + \hat{\sigma}_{u,m}^2}},$$

which are known as z-scores [13]. We ignore scores with $|z_{m,n}| > 2.5$ and workers with more than 5% of outlying scores. Note that this procedure is similar to the post-screening proposed in [3].

4. EXAMPLE: SCORING THE LIVE DATASET

Release 1 of the LIVE [14] dataset has ACR scores for 233 JPEG images and 227 JPEG 2000 images. For evaluating crowdMOS, we only used the JPEG subset, which was created by compressing 29 RGB images at different compression ratios. The dataset is provided with ACR scores obtained over two studies, which evaluated 116 and 117 images with 20 and 13 volunteers, respectively. The LIVE ACR scores were obtained under controlled conditions, and the volunteers were students enrolled in signal processing courses.

In the crowdMOS studies, workers scored 12 images per HIT, and could work until they rated all images. We paid the rewards and bonuses described in Section 2, which resulted in most users returning complete or nearly complete sets. We ran each study until every HIT was submitted by 40 workers, which required approximately 2 hours per study. The total budget for the JPEG LIVE dataset was approximately \$140. Note that this comes out to $\$140/233 \approx \$0.60/\text{image}$.

Fig. 2 shows MOS values and 95% CIs for the LIVE and crowdMOS results (due to space limitations, we only plot results for Study 1). The plotted values are z-scores recentered at 3.5 and scaled to fit the 1-5 scale (the official LIVE scores are z-scores stretched between 1-100).

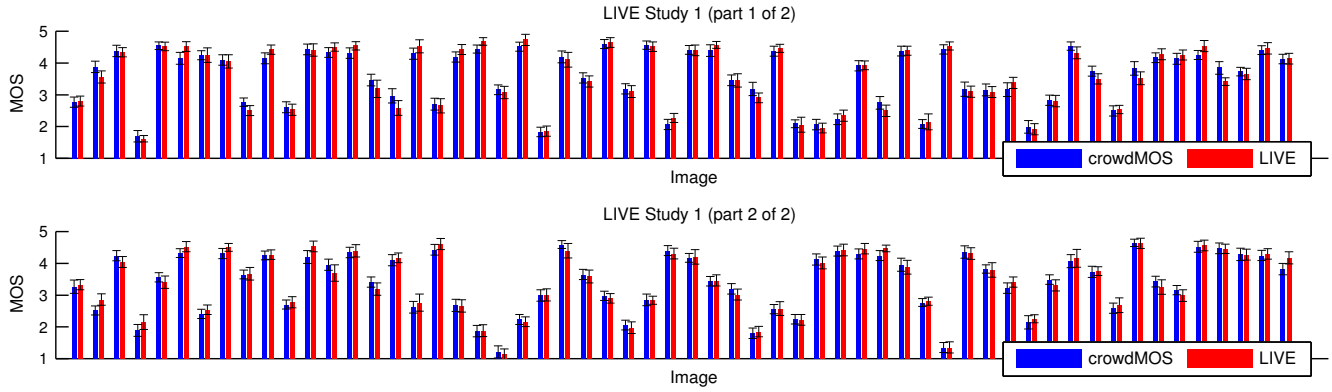

Fig. 2. Comparison between crowdMOS and LIVE laboratory scores, with 95% confidence intervals.

Table 2. Experiment sizes after outlier removal

Experiment	# of Scores	# of Workers
LIVE ACR Study 1	2187	19
LIVE ACR Study 2	3491	34
crowdMOS ACR Study 1	1510	13
crowdMOS ACR Study 2	3770	38

Table 3. Correlation coefficients with respect to LIVE scores

Experiment	r	ρ
crowdMOS ACR Study 1	0.985	0.965
crowdMOS ACR Study 2	0.981	0.948

Note that the unnormalized crowdMOS scores are more discriminating than the LIVE scores. Indeed, good or excellent images were scored approximately 0.5 MOS higher with crowdMOS. Bad or poor images were scored approximately 0.2 MOS lower. Analyzing individual scores, we noticed that the LIVE volunteers were very discriminating in identifying faint JPEG artifacts. They were also more tolerant of strong block effects. This suggests that their preferences and discrimination capacity were modified by their experience with signal processing. Nevertheless, this only appears to have linearly stretched their scales, since z-scores show very good agreement between crowdMOS and LIVE.

Table 2 shows the experiment sizes after outlier elimination. For the crowdMOS study, we only considered users who completed 5 or more HITs (thus, not all ignored workers were outliers). Table 3 shows the Pearson and Spearman rank order correlation coefficients (r and ρ, respectively) between the crowdMOS studies and the LIVE laboratory results.

5. CONCLUSION

In this paper we described crowdMOS, a family of subjective quality measures designed to enable subjective quality studies without the overhead and costs associated with laboratory testing. By combining crowdsourcing with user screening, we have shown that crowdMOS can deliver accurate, statistically significant results in very short timeframes and with low costs. Indeed, we ran full MOS studies in 2 hours at a cost of about $\$0.60$/image.

CrowdMOS applies to studies where impairment can be detected without high-end hardware, and expert training is not required. Thus, it can be used to complement or replace objective quality measures in preliminary quality assessments, where a rigorous MOS test is not yet justified, yet a large-scale subjective experiment is highly desirable. It has the characteristic (and potential advantage) of having random real world users with commodity hardware in their own environments. Because these are typical users, the difference between typical MOS studies conducted by researchers and crowdMOS are likely to highlight what is most important to real users.

One of the contributions of this work is an associated set of free, open-source software tools [8] designed to automate subjective quality experiments with Amazon Mechanical Turk. Using these tools, the technical and bookkeeping aspects surrounding crowdsourcing are hidden, and the researcher can concentrate on his experiment. We have briefly illustrated its use by running an ACR study with the JPEG images of the LIVE dataset, and obtaining a Pearson correlation coefficient of .98 with respect to the official laboratory results. Finally, the crowdMOS tools can be used out of the box to implement various types of standardized subjective tests, and can be easily modified to run other subjective experiments for signal processing, which are the subject of ongoing work.

6. REFERENCES

[1] "Subjective video quality assessment methods for multimedia applications," ITU-T Recommendation P.910, Apr. 2008.

[2] "Methods for subjective determination of transmission quality," ITU-T Recommendation P.800, Aug. 1996.

[3] "Methodology for the subjective assessment of the quality of television pictures," ITU-R Recommendation BT.500-12, Sept. 2009.

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

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

[6] D.M. Chandler and S.S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, 2007.

[7] A.K. Moorthy and A.C. Bovik, "Perceptually significant spatial pooling techniques for image quality assessment," in *Proc. of SPIE*, 2009.

[8] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "CrowdMOS Standalone Tools," available at http://research.microsoft.com/crowdmos/.

[9] K.T. Chen, C.C. Wu, Y.C. Chang, and C.L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proc. ACM Multimedia 2009*. ACM, 2009, pp. 491–500.

[10] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "CrowdMOS: an approach for crowdsourcing mean opinion score studies," in *Proc. of ICASSP*, 2011.

[11] "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," ITU-R Recommendation BS.1116-1, Oct. 1997.

[12] B. Efron, R. Tibshirani, and R.J. Tibshirani, *An introduction to the bootstrap*, Chapman & Hall/CRC, 1993.

[13] A.M. van Dijk, J.B. Martens, and A.B. Watson, "Quality assessment of coded images using numerical category scaling," in *Proc. of SPIE*, 1995.

[14] H.R. Sheikh, Z. Wang, L. Cormack, and A.C. Bovik, "LIVE image quality assessment database," 2003, http://live.ece.utexas.edu/research/quality/.