A PROBABILISTIC PAIRWISE-PREFERENCE PREDICTOR FOR IMAGE QUALITY

Amy R. Reibman, Kenneth Shirley, and Chao Tian
AT&T Labs – Research, Florham Park, NJ, USA

ABSTRACT

Current image quality estimators (QEs) compute a single score to estimate the perceived quality of a single input image. When comparing image quality between two images with such a QE, one only knows which image has a higher score; there is no knowledge about the uncertainty of these scores or what fraction of viewers might actually prefer the image with the lower score. In this paper, we present a Probabilistic Pairwise Preference Predictor (P^4) that estimates the probability that one image will be preferred by a random viewer relative to a second image. We train a multilevel Bayesian logistic regression model using results from a large-scale subjective test and present the degree to which various factors influence subjective quality. We demonstrate our model provides well-calibrated estimates of pairwise image preferences using a validation set comprising pairs with 60 reference images outside the training set.

1. INTRODUCTION

Systems for image capture, compression, transmission, and display can all benefit from a visual quality estimator (QE) that can accurately mimic subjective, human quality judgments for a wide range of input content and processing types. Unfortunately, obtaining accurate subjective quality estimates is expensive and time-consuming; perhaps even more problematic, human responses are inherently probabilistic. Inter-viewer variability may occur due to a viewer's individual ability to discriminate, a viewer's preference of one distortion type over another, or a preference for a distortion in one spatial region over another. Even the same viewer may produce different answers at different times, due to different viewing conditions, fatigue, lack of focus, or distractions. For example, greater inter-viewer variability is expected if the images being rated have similar quality or if the task is complex such as a preference among images with different content.

The question we seek to answer in this work is the following: can a QE accurately mimic human responses in a probabilistic manner? More precisely, we wish to find a pairwise-preference predictor which can provide an estimate of the form "a random viewer will prefer the image on the left with a probability of, e.g., 40%". Despite the growing recognition of viewer variability [1, 2], to our knowledge, no existing objective QE produces a probabilistic estimate.

We choose to model relative quality, e.g., "the image on the LEFT is better than that on the RIGHT", instead of absolute quality, e.g., "this image has a quality score of 4.5", for a variety of reasons. Firstly, relative QEs have many applications, such as product and algorithm benchmarking and selection [3]. Secondly, a relative QE allows for a probabilistic interpretation, which provides more context than an absolute QE (knowing that two images have absolute QE scores of 2.3 vs. 3.6, for example, is less informative, on its own, than knowing that 75% of viewers prefer one image to the other, which is interpretable even to a layperson). Lastly, subjective testing methods for absolute ratings require training (e.g., for viewers to recognize the meaning of the score and the correct usage of the dynamic range), and are usually very tightly prescribed [4]. This implies that data collection is limited, and the results thus obtained may not apply to real-life viewers and viewing conditions. On the other hand, when the question posed is *relative*, lightly trained viewers can complete the task competently [5]. Thus relative QEs are more suitable for large scale subjective tests and the probabilistic QE model we propose here.

Naively, one can use an existing QE to decide one image has a better (or worse) quality than another if their objective QE scores differ by more than a constant threshold Δo [6], and deem them to have equal quality otherwise. This is contingent upon the choice of an effective and meaningful Δo, and further does not reflect the probabilistic nature of viewer responses.

In this work, we present a probabilistic pairwise-preference predictor (P^4), which uses a Bayesian statistical model for pairwise preferences that is a function of properties of degraded images including reference image, distortion type, and an ensemble of QE scores. Alternate or additional QEs can be incorporated into the ensemble easily. However, in this study we choose six full-reference (FR) QEs: SSIM [7], IW-SSIM [8], PSNR-HVS-M [9], VIF [10], VSNR [11], and PSNR. In addition to creating a point estimate of the probability that one image is preferred relative to another, our model also provides an interval prediction based on its confidence in this point estimate.

Precisely because of the reduced training requirement and the need for large-scale data collection, we collect subjective data using Amazon Mechanical Turk (AMT). Prior studies that have explored using AMT [5, 12, 13, 14] or similar crowd-sourcing platforms [15, 16] have carefully considered worker reliability, consistency, and screening. The general consensus is that, given proper task formation, workers are usually sufficiently reliable. Furthermore, we believe the diversity of the viewers on AMT actually enables our model to accurately capture underlying viewer variation.

2. SUBJECTIVE TESTING

2.1. Background

The three largest publicly available subjective image quality datasets (see review article [17]) are the LIVE [18], TID2008 [19], and CSIQ [20] databases, which are summarized in Table 1. These subjective tests, using up to 30 reference images and up to 17 types of distortion, have been used a benchmark to evaluate the performance of QEs. These subjective datasets differ by the type of decision or label made by each viewer. Experiments requesting absolute scores from viewers typically request comparisons across different content [18], while many paired comparison experiments explicitly avoid such comparisons [19, 21, 5]. Nonetheless, all three report a subjective estimate of *absolute* visual quality for each image.

2.2. Our subjective tests

We begin by choosing a collection of reference, or source, images. The first 30 reference images are taken from the CSIQ database [20].

Name	# ref. images	# distortions	# HRC per ref img	total # distorted images
LIVE 2	29	5	26-28	779
CSIQ	30	6	28-29	866
TID2008	25	17	68	1700
RST90	90	4	118-119	10,690

Table 1. Overview of subjective tests

Class	Ref. images	Distortions	# pairs stage 1	# pairs stage 2
I	Same	Same	0	303
II	Same	Different	7601	4951
III	Different	Same	9492	5462
IV	Different	Different	0	5115

Table 2. Pair selection for our subjective test.

Each of the next 60 reference images was captured using a high-quality high-resolution digital camera from an outdoor scene. Each image is filtered, downsampled, and cropped to produce 512*512 pixels. Among these 60 images, there are 16 animal pictures, 17 landscapes, and 27 structures (including buildings and sculptures). The spatial information (SI) and colorfulness (CF) scores for each reference image are computed as described in [17], and indicate a slightly wider range of CF and SI than [18, 19, 20].

Next, we choose the four distortion types that appear in nearly all image quality databases: Gaussian blur, JPEG-2000 and JPEG compression, and additive Gaussian noise. For each distortion type, we choose 29-30 severity values, ranging from little distortion to moderately severe, which results in a total of 118-119 distorted images for each of our 90 reference images. We create pairs of images for a two-stage subjective test. In the first stage, we use only the 3550 distorted images obtained from the CSIQ images, while in the second stage we use distorted images from all 90 images with an emphasis on pairs taken from the 60 new reference images. The labeled pairs from the first stage are used to train our model, while the labeled pairs from the second stage are exclusively used to validate our model.

Pairs are constructed to emphasize important use cases of a QE. In particular, we decompose all possible pairs into four classes, based on whether both images share a common reference image or a common distortion type. As in [22, 23], our experiment contains comparisons across different reference images. Table 2 summarizes the number of pairs we choose in each class. In the second stage, 13860 pairs have only images from the new 60 reference images, while 1971 have one image from the 30 CSIQ reference images and the other image from one of our new reference images.

Using AMT, image pairs were presented in random order (and random left/right assignment), and the viewers, who were naive to the purposes of the experiment, were instructed to "click on the image with better visual QUALITY between the two images. Choose the image with the better technical quality, not the image content you prefer." Each task for a viewer contained 10 pairs, and each viewer was limited to a maximum number of 300 pairs. No image pair was rated by more than one viewer by design to obtain more efficient estimates of the effects of distortions. In total, 450 unique viewers participated in our study, and viewers whose data were clearly unreliable or showed extreme bias were rejected.

3. THE MODEL

We use logistic regression to model which image from a pair is chosen by a viewer. The model we fit is a multilevel Bradley-Terry model, where we model the latent subjective quality of each image as a function of (1) the reference image, (2) the distortion type applied to the image, and (3) six QEs, where the effect for each QE differs by distortion type. Modeling subjective quality as a function of image-level variables allows us to generalize our model to images outside those in our training data, so that we can make predictions about viewer preferences for new image pairs.

Let $Y_i = 1$ if the subject chose the left image in pair i, and $Y_i = 0$ if the subject chose the right image, for image pairs $i = 1, ..., N$, where in our training data, $N = 13,674$ (this represents 80% of the image pairs from stage 1 of our experiment, where the other 20% is held out for testing). Let $V[i]$ denote the viewer of image pair i, for viewers $w = 1, ..., W = 249$. Let $L[i]$ and $R[i]$ denote the reference image of the left and right image, respectively, in pair i, for reference images $j = 1, ..., J = 30$. Let Dist-L$[i]$ and Dist-R$[i]$ denote the distortion types applied to the left and right images in pair i, for distortion types $d = 1, ..., D = 4$. Last, let $X^{\text{QE-L}}_{k[i]}$ and $X^{\text{QE-R}}_{k[i]}$ be the objective quality score for the kth QE applied to the left and right images in pair i, respectively, for QEs $k = 1, \ldots, K = 6$. The QE scores are transformed (if necessary) and scaled to have a mean of zero and standard deviation of one so that their estimated effects are comparable; all increase monotonically with image quality.

The model is a multilevel (i.e. hierarchical) Bayesian logistic regression model:

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(p_i), \\
\log\left(\frac{p_i}{1-p_i}\right) &= \alpha^{\text{viewer}}_{V[i]} + \lambda^{\text{Left}}_i - \lambda^{\text{Right}}_i \qquad (1) \\
\lambda^{\text{Left}}_i &= \alpha^{\text{image}}_{L[i]} + \alpha^{\text{distortion-type}}_{\text{Dist-L}[i]} + \sum_{k=1}^{K} \beta^{\text{objective}}_{(k,\text{Dist-L})[i]} \times X^{\text{QE-L}}_{k[i]},
\end{aligned}
$$

for image pairs $i = 1, ..., N$, where λ^{Right}_i is defined analogously to λ^{Left}_i, and these represent the latent subjective qualities of the right and left images, respectively.

We use normal priors for the viewer, reference image, distortion-type, and QE effects:

$$
\begin{aligned}
\alpha^{\text{viewer}}_v &\sim \text{N}(\mu^{\text{viewer}}, \sigma^2_{\text{viewer}}), \\
\alpha^{\text{image}}_r &\sim \text{N}(0, \sigma^2_{\text{image}}), \\
\alpha^{\text{distortion-type}}_d &\sim \text{N}(0, \sigma^2_{\text{distortion-type}}), \\
\beta^{\text{objective}}_{kd} &\sim \text{N}(\mu^{\text{objective}}_k, \tau^2_{\text{objective-dist}_k}) \\
\mu^{\text{objective}}_k &\sim \text{N}(\mu_0, \tau^2_{\text{objective}}),
\end{aligned}
$$

We use weakly informative half-t priors for the standard deviation parameters σ_{viewer}, σ_{image}, $\sigma_{\text{distortion-type}}$, $\tau_{\text{objective-dist}_k}$ (for $k = 1, ..., 6$), and $\tau_{\text{objective}}$, and $\text{N}(0, 1)$ priors for μ^{viewer} and μ_0 [24].

We fit the model using the the Bayesian MCMC software package, JAGS ("Just Another Gibbs Sampler") [25]. The model converges almost immediately, although it takes about 70 minutes to sample 5000 iterations from the posterior distribution of all 327 unknown parameters. We discarded the first 1000 iterations as burn-in, and kept every 10th iteration from the next 4000 iterations as our posterior sample, for each of 3 independent chains, giving us a total of 1200 posterior samples, each with 327 parameters.

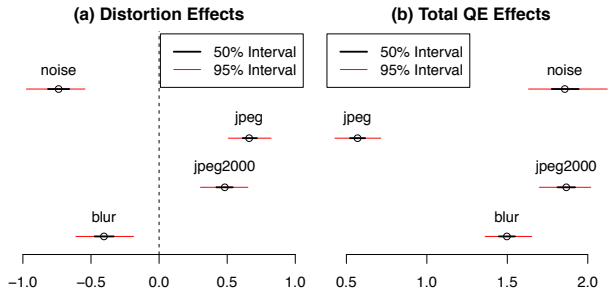

Fig. 1. Interval estimates of $\alpha^{\text{distortion}}$ and $\beta^{\text{obj-sum}}$

4. INFERENCE AND GOODNESS-OF-FIT

First, we describe point estimates and interval estimates for various parameters in our model. The mean left/right bias in the population of subjects, μ_{viewer}, was about 0.06 (on the logistic scale). This corresponds to a probability of a random viewer picking the left image with probability 51.5%. The viewer bias effect, however, was small compared to the effects of the other factors in the model. The estimated standard deviations for the viewer, reference image, and distortion type effects were $\sigma_{\text{viewer}} \approx 0.19\ (0.04)$, $\sigma_{\text{image}} \approx 0.44\ (0.07)$, and $\sigma_{\text{distortion}} \approx 1.03\ (0.60)$, where standard errors are listed in parentheses. In other words, of these three factors, the distortion type explained the most variation in the outcome, and the viewer bias explained the least variation. To interpret these group-level standard deviations, consider that on the logistic scale, holding all other variables at their observed values, choosing two different distortion types at random would induce an expected change in the probability of choosing the left image of about 25% – a large effect. Randomly choosing two reference images, or two viewers, would affect the probability of choosing the left image by about 12% or 5%, on average, respectively.

Regarding the most extreme variations, the most biased viewers in our experiment had approximately a 44% and 59% probability of choosing the left image, holding all other variables constant. The most preferred reference image, all else held constant, was "sunset-color" (74.0% chance of being preferred compared to the average reference image), and the least preferred was "fisher" (26.7%).

The estimated effects for distortion types and QEs are pictured in Figure 1. Figure 1(a) shows that JPEG and JPEG2000 distortions are preferred over Blur and Noise distortions across the images in our collection, where a JPEG-distorted image would be preferred over a noise-distorted image (all else held constant) about 80% of the time, for example. The 24 objective quality effects, $\beta_{kd}^{\text{objective}}$ (for $k = 1, ..., 6$ and $d = 1, ..., 4$), are more complicated to interpret, since all six QEs are highly correlated with each other. To summarize their effects, we estimate the posterior distribution of the *sum* of their effects for each distortion type. Define the "Total QE effect" for each distortion type d as $\beta_d^{\text{obj-sum}} = \Sigma_{k=1}^{K} \beta_{kd}^{\text{objective}}$. The estimates of these sums are precise – they are 1.50, 1.86, 0.57, and 1.86 for the distortion types Blur, JPEG2000, JPEG, and Noise, respectively, with standard errors less than 0.12 in all four cases (pictured in Figure 1(b)). This means that this collection of six QEs has the strongest association with subjective quality for JPEG2000 and Noise distortion types, and the weakest association for JPEG distortions.

Second, to check the fit of the model, we make predictions on the holdout set, which consists of 20% of the pairs in our data from Stage 1. For each image pair in the holdout set, and for each posterior

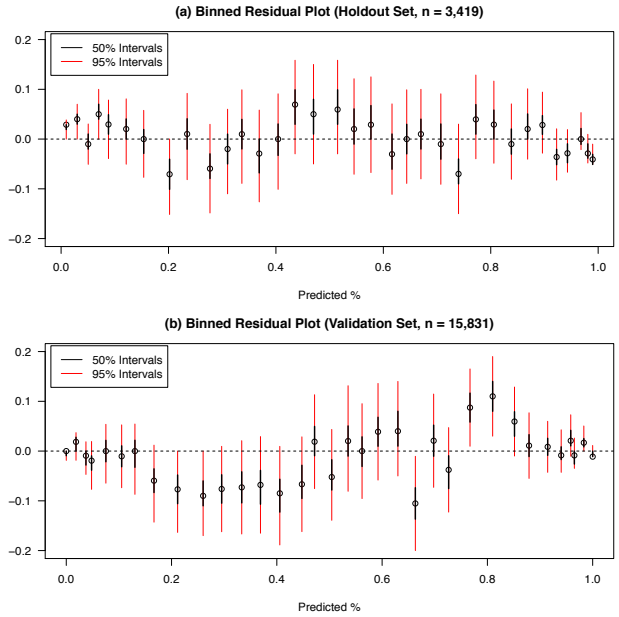

Fig. 2. (a) Binned residual plot for the holdout set; RMSE is 3.5%. (b) Binned residual plot for the validation set; RMSE is 4.9%, and there is a slight pattern of shrinkage in the residuals. $n_{\text{bin}} = 100$

sample of parameters, we compute the estimated probability that the viewer chose the left image, and we then simulate a response from the Bernoulli distribution with this probability. From these simulations, we draw binned residual plots to look for lack of fit in our model. Figure 2(a) contains a binned residual plot where the data points (pairs) are binned by their posterior mean probability of the viewer choosing the left image. The differences between the actual proportion of viewers who chose left in each bin and the predicted proportions are centered around zero, with no discernible pattern, and the 50% and 95% intervals have the advertised coverage. This is a strong indication that the model fits well with respect to data that comes from the same population as the data to which the model was fit. Note, however, that when making these predictions, we knew the identities of the viewers of the holdout pairs, we knew the reference images, and we knew the distortion types; this is unlikely to be true for paired-comparison predictions "in the wild". We discuss predictions for new image pairs in Section 5.

Third, knowing that the model fits the data, the last question remaining is, "How accurate is the model?". We recommend two easy-to-interpret measures of predictive accuracy for our model. First, we measure the Root Mean Squared Error (RMSE) of our model's predictions of the percentage of pairs in which the viewer prefers the left image, for some standard bin size. Here, we choose $n_{\text{bin}} = 100$ pairs. In Figure 2(a), this is simply the RMSE of the differences between the black points and the horizontal line at zero. For our holdout data, the RMSE for $n_{\text{bin}} = 100$ is 3.5%, and the errors are approximately normally distributed. This means that about 2/3 of the time our model's prediction will be within 3.5% of the true percentage. We also compute the misclassification error, where we use the posterior mean of $\text{P}(Y_i^{\text{holdout}} = 1)$ to classify each pair as either having its left or right image chosen by the viewer. The misclassification rate was 22.8% for the pairs in our holdout set. For comparison, in a sample of 400 image pairs labeled by two experts, the experts disagreed on 16% of image pairs – providing a rough gold standard for

the performance of any statistical model fit to this data.

5. APPLYING THE MODEL TO NEW DATA

In this section, we describe how to use our model to make predictions for new image pairs outside the training population. The basic principle is simple: in the absence of knowing the effect of a particular variable (say, the viewer effect), we must sample an effect from the *distribution* of that variable's effects (i.e. from the $\text{N}(\mu^{\text{viewer}}, \sigma^2_{\text{viewer}})$ distribution). This propagates our uncertainty of a given variable's effect through the model into our predictions.

The precise steps are as follows. For each posterior sample $g = 1, ..., G$ (where $G = 1200$ in this paper), let $\theta^{(g)}$ denotes the gth sample of the parameter θ in the set of posterior samples, and let L,R denote the Left and Right images in the new pair, respectively. If we don't know the viewer, reference image, or distortion type for either L or R:

1. Draw a viewer effect from $\text{N}(\mu^{(g)}_{\text{viewer}}, \sigma^{(g)}_{\text{viewer}})$.
2. Draw a reference image effect for L from $\text{N}(0, \sigma^{(g)}_{\text{image}})$.
3. Draw a distortion type effect for L from $\text{N}(0, \sigma^{(g)}_{\text{distortion-type}})$.
4. Draw a QE effect for the kth QE score for L from $\text{N}(\mu^{\text{objective (g)}}_k, \tau^{(g)}_{\text{objective-dist}_k})$, for $k = 1, ..., 6$, since we assume that we don't know the distortion type d for L.
5. Repeat steps 2-4 for R, unless it is known that R and L share a common reference image or distortion type. In those cases, set the effect for R equal to that drawn for L.
6. Compute the estimate of $\text{P}(Y^{\text{new}}_i = 1)$ from Equation 1.

Now we have the posterior distribution of the probability of choosing the left image for each new image pair, which is wider (i.e. larger standard deviation) because now we don't know the viewer, or reference image, or distortion type. When either the reference images or the distortion types are known to be the same, then the posterior distribution is narrower, but not as narrow as when all variables are known.

For an illustration of making a prediction for a new image pair, see Figure 3. In this example, we look at an image pair from the holdout set in which reference image 10 ("family") is shown on the left, reference image 2 ("aerial_city") is shown on the right, the viewer $w = 74$ (from the training data), the distortion type is noise for each image, and the six QE scores are known for each image. If we make a prediction for this image pair using information from training, namely, that $\hat{\alpha}^{\text{viewer}}_{74} \approx 0.01$, $\hat{\alpha}^{\text{image}}_{10} \approx -0.05$, and $\hat{\alpha}^{\text{image}}_{2} \approx -0.10$, and factoring in the QE effects, we estimate the posterior mean of the probability of choosing the left image is 0.35, with a standard error of about 0.06. If this image pair were new, however, and we hadn't known the viewer, and the reference images were new and different from each other, and if we assumed that we knew the distortion types were the same (allowing these values to cancel), we would estimate the probability of choosing the left image to be 0.38 with a standard error of 0.19. In other words, the probability would move toward 50%, and our confidence in the probability estimate would be lower.

Next we check the accuracy of our predictions for new viewers and new reference images by making predictions on our validation set (Stage 2 of our experiments), which contains $N^{\text{validate}} = 15,831$ image pairs, with new viewers and new reference images not in the training data. Following the procedure outlined above, we computed $G = 1200$ samples from the posterior distributions of the probability of choosing the left image in each of the validation image pairs.

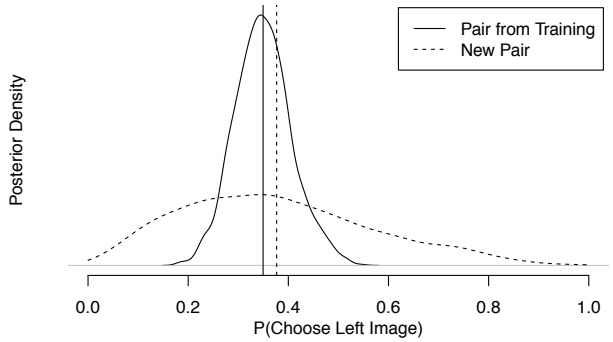

Fig. 3. Density estimates for the posterior distribution of $\text{P}(Y^{\text{holdout}}_i = 1)$ given that we (a) know the viewer, distortion types, and reference images, and (b) know none of these variables (although we assume unequal reference images and equal distortion types to match the original training data).

Among the 246 viewers who participated in Stage 2, 45 of them had also participated in Stage 1, such that we had estimated their viewer effects from our training data. Also, some of the images in the validation data had reference images in common with images from Stage 1. For pairs with known viewer effects or reference image effects, we used the estimated effects from training, but for new viewers and reference images, we followed the outlined procedure.

Figure 2(b) contains a binned residual plot for predictions made on the validation set. We use bin size $n_{\text{bin}} = 100$, but show only 34 bins in Figure 2(b), to create a visual comparison to the analysis on the holdout set in Figure 2(a). The RMSE of the predicted percentages is 4.9%, and the pattern of residuals indicates that the model is "shrinking" estimated probabilities slightly too far toward 50% for pairs in which the predicted probability is between 20% and 40%. This could be a sign that the variation between viewers or reference images in Stage 2 is *less* than that of Stage 1, and hence the estimates of σ_{viewer} and σ_{image} from the training set were too large, and induced extra randomness in predictions. The misclassification rate for the validation set was 19.6%, substantially lower than that of the holdout set. If we look at misclassification rate with respect to the four classes of pairs described in Table 2, we see slight variation in the results: the rates for each class are 19.1%, 24.2%, 15.2%, and 17.8%, for classes I - IV, respectively. Recall that the model was trained only using points from Classes II and III. Further experiments can be conducted to improve the fit and accuracy of the model, and shed light on whether the effects of new distortion types are accurately predicted by our model.

6. CONCLUDING THOUGHTS

We propose a probabilistic pairwise-preference predictor (P^4), which estimates the probability that a given image in a pair will be preferred by a random viewer, and includes interval estimates to gauge prediction uncertainty. Our model is by no means complete, and a more complex model may need to incorporate further interactions, such as interactions among different QEs. Although only four types of distortions were included in this study, our approach can be further generalized to include other common distortion types. In fact, due to the robustness of the data collection process for the pairwise preference subjective tests, the data we collected for this work can also be reused and incorporated into a future data collection effort with other types of distortions.

7. REFERENCES

[1] U. Engelke, Y. Pitrey, and P. Le Callet, "Towards a framework of inter-observer analysis in multimedia quality assessment," *QoMEX*, 2011.

[2] E. Karapanos, J.-B. Martens, and M. Hassenzahl, "Accounting for diversity in subjective judgements," in *ACM Computer Human Interface*, Apr. 2009.

[3] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Processing: Image Communication*, Aug. 2010.

[4] ITU-R Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva, Switzerland, Jan. 2012.

[5] K.T. Chen, C.C. Wu, Y.C. Chang, and C.L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 491–500.

[6] International Telecommunication Union, "J.149: Method for specifying accuracy and cross-calibration of video quality metrics (VQM)," Mar. 2004.

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

[8] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Proc.*, vol. 0, no. 5, pp. 1185–1198, May 2011.

[9] N. Ponomarenko et al. "On between-coefficient contrast masking of DCT basis functions," in *Wkshp. on Video Proc. and Quality Metrics*, 2007.

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

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

[12] F. Cole, K. Sanik, D. DeCarlo, A. Finkelstein, T. Funkhouser, S. Rusinkiewicz, and M. Singh, "How well do line drawings depict shape?," *ACM Trans. Graphics*, vol. 28, no. 3, pp. 28:1–28:9, July 2009.

[13] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," *ACM Trans. Graphics*, vol. 29, no. 6, pp. 160:1–160:10, Dec. 2010.

[14] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *IEEE Int. Conf. Image Proc.*, 2011.

[15] C. Keimel, J. Habigt, and K. Diepold, "Challenges in crowd-based video quality assessment," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, 2012, pp. 13–18.

[16] D.R. Rasmussen, "The mobile image quality survey game," in *SPIE Image Quality and System Performance IX*, 2012, vol. 8293, p. 16.

[17] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616 –625, Oct. 2012.

[18] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Proc.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[19] N. Ponomarenko et al., "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 2009.

[20] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. of Electronic Imaging*, vol. 19, no. 1, Mar. 2010, http://vision.okstate.edu/index.php?loc=csiq.

[21] F. M. Ciaramello and A. R. Reibman, "Supplemental subjective testing to evaluate the performance of image and video quality estimators," in *Human Vision and Electronic Imaging XVI*, Jan. 2011.

[22] G.O. Pinto and S.S. Hemami, "Image quality assessment in the low quality regime," in *IS&T/SPIE Electronic Imaging*, 2012, vol. 8291.

[23] A. R. Reibman, "A strategy to jointly test image quality estimators subjectively," in *IEEE Int. Conf. Image Proc.*, Sept. 2012.

[24] A. Gelman, A. Jakulin, M. G. Pittau, and Y. S. Su, "A weakly informative default prior distribution for logistic and other regression models," *Annals of Applied Statistics*, vol. 2, no. 4, pp. 1360–1383, 2008.

[25] M. Plummer, "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," in *3rd International Workshop on Distributed Statistical Computing*, 2003.