Active Sampling for Subjective Image Quality Assessment

Peng Ye and David Doermann
Institute for Advanced Computer Studies
University of Maryland, College Park, MD, USA
{pengye,doermann}@umiacs.umd.edu

Abstract

Subjective Image Quality Assessment (IQA) is the most reliable way to evaluate the visual quality of digital images perceived by the end user. It is often used to construct image quality datasets and provide the groundtruth for building and evaluating objective quality measures. Subjective tests based on the Mean Opinion Score (MOS) have been widely used in previous studies, but have many known problems such as an ambiguous scale definition and dissimilar interpretations of the scale among subjects. To overcome these limitations, Paired Comparison (PC) tests have been proposed as an alternative and are expected to yield more reliable results. However, PC tests can be expensive and time consuming, since for n images they require $\binom{n}{2}$ comparisons. We present a hybrid subjective test which combines MOS and PC tests via a unified probabilistic model and an active sampling method. The proposed method actively constructs a set of queries consisting of MOS and PC tests based on the expected information gain provided by each test and can effectively reduce the number of tests required for achieving a target accuracy. Our method can be used in conventional laboratory studies as well as crowdsourcing experiments. Experimental results show the proposed method outperforms state-of-the-art subjective IQA tests in a crowdsourced setting.

1. Introduction

Estimating gold-standard labels (strengths, scores, etc.) based on subjective judgments provided by humans is a critical step in psychological experiments with applications in many research fields [19]. This paper studies the problem of Quality of Experience (QoE) evaluation, which in general aims to obtain subjective satisfaction of user's experience with a service (for example, web browsing, phone calls, video chatting or online shopping.) or with some multimedia content (for example, videos, images, etc.). Here we investigate image quality assessment (IQA) problem, but the proposed method can be applied to any general problem of QoE evaluation.

Absolute Category Rating [1] is one of the most popular subjective IQA tests. It consists of having a panel of subjects rate images using an ordinal scale: 1-"Bad", 2-"Poor", 3-"Fair", 4-"Good" and 5-"Excellent". For a given image, its score is computed as the average scores from all subjects. This is also known as the Mean Opinion Score (MOS). Despite the popularity of the MOS test, there are many known problems [4, 5]. First, most previous work in QoE [16] treats the MOS scale as an interval scale instead of ordinal scale and assumes that the cognitive distances between the consecutive MOS scales are the same. However, assumptions such as: "Fair"-"Poor"="Good"-"Fair", are not always true in practice. Second, absolute rating procedures are somewhat obscure so subjects can be easily confused about which scale they should give in each test and different subjects may have different interpretations of the scale. Therefore the resulting rating observations can be very noisy.

To overcome the limitation of the MOS test, the Paired Comparison (PC) test [4, 5, 8, 18, 19, 23, 24] has been proposed as an alternative. In the simplest configuration, two images *A* and *B* are shown to a subject who is asked to "prefer" one of them. Compared to the rating test, making a decision in a paired comparison test is much simpler and less confusing for the subject. However, when n images need to be compared, the total number of pairs is $\binom{n}{2}$ and when n is large, the cost for obtaining a full set of pairwise comparisons is prohibitively expensive. HodgeRank on Random Graphs (HRRG) [23, 24] has been introduced to reduce the cost of the PC test by using a random sampling method with HodgeRank [11]. We approach this problem differently by combining the MOS test and the PC test via active sampling. As will be shown experimentally, the proposed method outperforms the HRRG for crowdsourcing subjective IQA.

Our method is motivated by the following observations: 1) Although the MOS test may not be able to accurately rank two images with similar quality due to the observation noise, it can provide an estimate of the underlying quality

1063-6919/14 $31.00 © 2014 IEEE
DOI 10.1109/CVPR.2014.541

IEEE computer society

score at a coarse level. 2) In the PC test, we explicitly ask humans to compare pairs of images, therefore the PC test can provide fine discrimination on images with similar quality. 3) Once we have some coarse estimates on the underlying scores, a complete set of PC test would be unnecessary. For example, it would be unnecessary to perform a paired comparison on an image with MOS score 1 and an image with MOS score 5, since we can already tell the difference with high confidence. Based on these observations, we will show that combining the MOS and PC tests will provide a more efficient design for subjective IQA. In this paper, we will answer the following two questions:

1. Given a collection of observations from the MOS test and the PC test, how can we combine them to estimate the underlying score?

2. In both laboratory studies and crowdsourced settings, subjective judgments are obtained at a defined cost. How can we effectively sample a subset of MOS and PC tests so that we can achieve desired accuracy with minimal cost?

2. Related Work

2.1. Crowdsourceable QoE

Conventional subjective QoE experiments conducted in laboratory settings can be expensive and time-consuming and typically only a small number of subjects are involved. With the ubiquitous internet access and the rise of internet micro-labor markets such as Amazon Mechanical Turk, there has been an increasing interest in designing subjective QoE tests for crowdsourced settings.

Previous work on Crowdsourceable QoE considers the MOS and PC tests independently. Ribeiro et al. [16] performed the MOS test for QoE assessment using crowdsourcing. They developed a two-way random effects model to model the uncertainty in subjective tests and proposed a post-screening method and rewarding mechanism to facilitate the process. Chen et al. [5] proposed a crowdsourceable QoE assessment framework for multimedia content, in which interval-scale scores are derived from a full set of paired comparisons. However, since a complete set of paired comparisons has to be performed, this method cannot be applied on a large scale. To address this problem, Xu et al. [23, 24] introduced the HodgeRank on Random Graphs (HRRG) test, where random sampling methods based on Erdös-Rényi random graphs were used to sample pairs and the HodgeRank [11] was used to recover the underlying quality scores from the incomplete and imbalanced set of paired comparisons. This method can effectively reduce the cost of PC tests required for achieving a certain accuracy. We will show experimentally that by combining information from MOS and PC tests via active sampling, we can further reduce the cost of experiments.

2.2. Preference Aggregation

The problem we are trying to solve is essentially an information aggregation problem, where we want to integrate information from multiple sources into a consensus score. The problem of preference aggregation has been extensively studied in the information retrieval community [6, 14, 20]. In particular, there has been some recent work in this field that applied active learning for preference aggregation. Given a pair of objects, a utility function is defined that measures the "usefulness" of performing a paired comparison. Then pairs with high utilities are chosen as queries and sent to an oracle or to human subjects.

Pfeiffer et al. [14] introduced an active learning method based on the Thurstone-Mosteller model [13, 18] for pairwise rank aggregation. At each iteration of an experiment, this method adaptively chooses one pair of objects to compare. The paper shows the advantage of using an active sampling method over a random sampling method. Chen et al. [6] proposed an active learning model based on the Bradley-Terry Model [3] which adopts an efficient online Bayesian updating scheme that does not require retraining of the whole model when new observations are obtained. All these previous works focus solely on aggregating information obtained from PC tests. A single optimal pair is usually chosen at each iteration of the experiment. This is inefficient in a crowdcourced setting, where multiple subjects may work in parallel and workers may expect to work on multiple tests instead of taking one single test in each working session. It is desirable to develop a batch-mode active learning method for the crowdsourceable subjective QoE problem.

Gleich and Lim [10] introduced several ad-hoc methods for building a preference matrix from rating observations based on the arithmetic mean of score differences, geometric mean of score ratios, binary comparisons, strict binary comparisons and logarithmic odds ratios. We may apply these methods to convert the rating observations into the preference observations. However, it is not clear how to measure the utility of the MOS test. Our method combines the MOS test and the PC test directly via a unified probabilistic model and the utility of each individual MOS test and PC test is defined as the expected information gain given by the test.

3. Combining Ratings and Paired Comparisons

This section presents a probabilistic model for combining the MOS test and the PC test. Suppose we have n images $A_1, A_2, .., A_n$ with underlying scores $s = (s_1, s_2, ..., s_n)$. We model a subject's perceived quality of image A_i as a random variable: $r_i = s_i + \varepsilon_i$, where the noise term is a Gaussian random variable $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

In the remainder of this section, we first derive the likelihood functions of the underlying score given the MOS and PC observations independently. We then present a hybrid system which estimates the underlying score using Maximum A Posteriori estimation (MAP).

3.1. Mean Opinion Score Test

Thurstone's law of categorical judgment [19] is applied for analyzing the rating observations. Assume the perceived categorical observation for A_i is m_i and $m_i \in \mathcal{M}$, where $\mathcal{M}$ is a finite set of K ordered categories and $K = 5$ in the case of the MOS test. Without loss of generality, these categories are denoted as consecutive integers: $\mathcal{M} = \{1, 2, ..., K\}$. We further introduce a set of cutoff values $-\infty \equiv \gamma_0 < \gamma_1 < \gamma_2 < ... < \gamma_{K-1} < \gamma_K \equiv \infty$[1]. When r_i falls between the cutoffs γ_{c-1} and γ_c, the observed categorical label is c, i.e. $m_i = c$, and we have

$$\begin{aligned} &Pr(m_i|s_i) = Pr(\gamma_{m_i-1} < s_i + \varepsilon_i \leq \gamma_{m_i}) \\ &= \Phi(\tfrac{\gamma_{m_i}-s_i}{\sigma}) - \Phi(\tfrac{\gamma_{m_i-1}-s_i}{\sigma}) \end{aligned} \tag{1}$$

where $\Phi(\cdot)$ represents Cumulative Density Function (CDF) of standard Gaussian distribution.

In the MOS test, repeated observations are made for each image. We define the rating observation matrix M as follows:

$$M = \begin{pmatrix} M_{1,1} & M_{2,1} & \cdots & M_{n,1} \\ M_{1,2} & M_{2,2} & \cdots & M_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ M_{1,K} & M_{2,K} & \cdots & M_{n,K} \end{pmatrix} \tag{2}$$

where $M_{i,j}$ is the number of times the image A_i is observed as in the j-th category. Given the underlying score s, we assume the categorical observations of each image are conditionally independent and follow a multinomial distribution. We then have the probability of observing M as follows:

$$\begin{aligned} &Pr(M|s) = \textstyle\prod_{i=1}^{n} Pr(M_{i,1}, M_{i,2}, ..., M_{i,K}|s_i) \\ &= \textstyle\prod_{i=1}^{n} \binom{M_{i,1}+...+M_{i,K}}{M_{i,1},...,M_{i,K}} \prod_{k=1}^{K} Pr(m_i = k|s_i)^{M_{i,k}} \\ &= c_1 \textstyle\prod_{i=1}^{n} \prod_{k=1}^{K} (\Phi(\tfrac{\gamma_k - s_i}{\sigma}) - \Phi(\tfrac{\gamma_{k-1}-s_i}{\sigma}))^{M_{i,k}} \end{aligned} \tag{3}$$

where c_1 is a constant.

3.2. Paired Comparison Test

In the PC test, if the perceived score $r_i > r_j$, we say that A_i is preferred to A_j, which is denoted as $A_i \succ A_j$. The probability of $A_i \succ A_j$ is given by:

$$Pr(A_i \succ A_j) = Pr(s_i + \varepsilon_i > s_j + \varepsilon_j) = \Phi(\tfrac{s_i - s_j}{\sqrt{2}\sigma}) \tag{4}$$

Eq. 4 is known as the Thurstone-Mosteller Case V model [13, 18]. Preferences obtained from a set of PC tests can be characterized by a preference matrix and we define the preference matrix P as:

$$P = \begin{pmatrix} 0 & P_{1,2} & \cdots & P_{1,n} \\ P_{2,1} & 0 & \cdots & P_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n,1} & P_{n,2} & \cdots & 0 \end{pmatrix} \tag{5}$$

where $P_{i,j}$ is the number of times $A_i \succ A_j$ is observed. Then the probability of observing P is:

$$\begin{aligned} &Pr(P|s) = \textstyle\prod_{i,j \in 1,...,n, i<j} Pr(P_{i,j}, P_{j,i}|s_i, s_j) \\ &= \textstyle\prod_{i<j} \binom{P_{i,j}+P_{j,i}}{P_{i,j}} Pr(A_i \succ A_j)^{P_{i,j}} Pr(A_j \succ A_i)^{P_{j,i}} \\ &= \textstyle\prod_{i<j} \binom{P_{i,j}+P_{j,i}}{P_{i,j}} \Phi(\tfrac{s_i - s_j}{\sqrt{2}\sigma})^{P_{i,j}} \Phi(\tfrac{s_j - s_i}{\sqrt{2}\sigma})^{P_{j,i}} \\ &= c_2 \textstyle\prod_{i \neq j} \Phi(\tfrac{s_i - s_j}{\sqrt{2}\sigma})^{P_{i,j}} \end{aligned} \tag{6}$$

where c_2 is a constant.

3.3. Posterior Probability of the Underlying Score

Given observations from both MOS and PC tests, the hybrid system estimates the underlying score by maximizing the posterior probability

$$\hat{s} = argmax_s Pr(s|P, M) \tag{7}$$

Computing $Pr(s|P, M)$ is not a trivial task. The likelihood functions in Eq. 3 and Eq. 6 are conditioned on several unknown model parameters including: the noise variance σ and the cut-off parameters $\gamma_1, ..., \gamma_{K-1}$. Since the likelihood functions are scale-invariant, i.e. $Pr(M|s, \gamma, \sigma) = Pr(M|ks, k\gamma, k\sigma)$ and $Pr(P|s, \sigma) = Pr(P|ks, k\sigma)$ for a constant $k \neq 0$, without loss of generality, we may fix $\sigma = 1/\sqrt{2}$. With σ fixed, the likelihood functions are still translation-invariant, i.e. $Pr(M|s, \gamma) = Pr(M|s + k, \gamma + k)$ and $Pr(P|s) = Pr(P|s + k)$ for a constant k. To make the objective identifiable, we further assume $\gamma_1 = 0$. $K - 2$ model parameters $\gamma_2, ..., \gamma_{K-1}$ remain unknown. We denote the set of unknown model parameters $\gamma = \{\gamma_2, ..., \gamma_{K-1}\}$.

In a full Bayesian treatment, computing $Pr(s|P, M)$ requires integrating the model parameters over all possible values, which can be implemented using Monte Carlo methods. However, these computations might be prohibitively expensive. Alternatively, we approximate $Pr(s|P, M)$ by $Pr(s|P, M, \hat{\gamma})$ where $\hat{\gamma}$ refers to the optimal setting of

[1] The original law of categorical judgment [19] assumes the randomness of γ_c, for simplicity, we assumes γ_c to be deterministic as in [7].

γ. Specifically, $\hat{\gamma} = argmax_{\gamma} Pr(M, P|\gamma)$, which is the Maximum Likelihood Estimate of γ. To obtain an analytical form of the gradients of $Pr(M, P|\gamma)$ w.r.t γ and a Gaussian form approximation to the posterior probability $Pr(s|M, P, \hat{\gamma})$, we apply the Laplace approximation [2]. To illustrate the approximation procedure, let us define:

$$\mathcal{F}_{\gamma}(s) = -logPr(M|s,\gamma) - logPr(P|s) - logPr(s) \quad (8)$$

where we assume a Gaussian prior on $s \sim N(\mu, \Omega)$. The Hessian matrix of $\mathcal{F}_{\gamma}(s)$ is given by:

$$R_{\gamma}(s) = \frac{\partial^2 \mathcal{F}_{\gamma}(s)}{\partial s \partial s^T} \quad (9)$$

Denoting the minimizer of $\mathcal{F}_{\gamma}(s)$ as $\hat{s}_{\gamma}$ and $\hat{R}_{\gamma} = R_{\gamma}(\hat{s}_{\gamma})$, applying a Laplace approximation, we have

$$\mathcal{F}_{\gamma}(s) \approx \mathcal{F}_{\gamma}(\hat{s}_{\gamma}) + \frac{1}{2}(s - \hat{s}_{\gamma})^T \hat{R}_{\gamma}(s - \hat{s}_{\gamma}) \quad (10)$$

Using the above approximation, $Pr(M, P|\gamma)$ can be computed analytically as follows:

$$\begin{aligned} &Pr(M, P|\gamma) = \int Pr(s)Pr(M|s,\gamma)Pr(P|s)ds \\ &= \int exp(-\mathcal{F}_{\gamma}(s))ds \approx exp(-\mathcal{F}_{\gamma}(\hat{s}_{\gamma}))(2\pi)^{\frac{n}{2}}|\hat{R}_{\gamma}|^{-1/2} \end{aligned} \quad (11)$$

Using Eq. 11, the gradients of the $log(Pr(M, P|\gamma))$ w.r.t γ can be computed analytically. Gradient-based optimization methods can be used to find MLE of γ.

Given the optimal cut-off parameter $\hat{\gamma}$, the posterior probability of s can be approximated by:

$$\begin{aligned} &Pr(s|P, M) \propto Pr(M|s, \hat{\gamma})Pr(P|s)Pr(s) \\ &= exp(-\mathcal{F}_{\hat{\gamma}}(s)) \propto N(\hat{s}_{\hat{\gamma}}, \hat{R}_{\hat{\gamma}}^{-1}) \end{aligned} \quad (12)$$

The MAP estimate of s is $\hat{s}_{\hat{\gamma}}$. In order to ensure a global optimal solution of the MAP estimate, Eq. 8 has to be a convex function. It has been shown in [7] that $-logPr(M|s,\gamma) - log(Pr(s))$ is convex. However, in order to make sure $-log(Pr(P|s))$ has a unique minimizer, Ford's condition [9] has to be satisfied. In practice, this can be achieved by adding a small constant to each zero-valued element in the preference matrix P. This is also known as smoothing.

4. Active Sampling

Subjective judgments are usually obtained at certain cost and it is desirable to design cost-efficient experiments. We propose an active random sampling method which constructs a set of queries consisting of MOS and PC tests based on the expected information gain provided by each. Let $\mathcal{E}_i$ denote the experiment which makes one absolute judgment on the object A_j and $\mathcal{E}_{ij}$ be the experiment that makes a pairwise comparison between A_i and A_j.

4.1. Information Measure of Experiments

The purpose of experiments is to gain knowledge about the state of nature. We adopt the Bayesian Optimal Design framework introduced by Lindley [12] and evaluate an experiment using the Expected Information Gain (EIG) provided by conducting this particular experiment. In the subjective IQA problem, the state of nature (or parameter) to be estimated is the quality score $s = \{s_1, ..., s_n\}$. Before conducting the experiment $\mathcal{E}$, our knowledge of s is characterized by the prior distribution of $s \sim Pr(s)$. The EIG provided by an experiment $\mathcal{E}$ is denoted $I(\mathcal{E}, Pr(s))$. The general formula of $I(\mathcal{E}, Pr(s))$ is given by [12]:

$$I(\mathcal{E}, Pr(s)) = E_s\left[\int log\{\frac{Pr(x|s)}{Pr(x)}\}Pr(x|s)dx\right] \quad (13)$$

where $E_s(\cdot)$ is the expectation taken w.r.t $Pr(s)$. For the MOS test, suppose the outcome of $\mathcal{E}_i$ is $x_i \in \{1, 2, ..., K\}$ and $p_{ik} = P(x_i = k|s)$. It is easy to verify that $p(x_i = k) = E_s(p(x_i = k|s)) = E_s(p_{ik})$ and we have

$$\begin{aligned} &I(\mathcal{E}_i, Pr(s)) = E_s[\textstyle\sum_{k=1}^{K} p_{ik} log(\frac{p_{ik}}{p(x_i=k)})] \\ &= E_s[\textstyle\sum_{k=1}^{K} p_{ik} log(p_{ik})] - \textstyle\sum_{k=1}^{K} E_s(p_{ik}) log E_s(p_{ik}) \end{aligned} \quad (14)$$

For the PC test, suppose the outcome of $\mathcal{E}_{ij}$ is x_{ij} and $x_{ij} = 1$ if $A_i \succ A_j$; $x_{ij} = 0$ if $A_i \prec A_j$. Define $p_{ij} = p(x_{ij} = 1|s)$ and $q_{ij} = 1 - p_{ij}$. It is easy to verify that $p(x_{ij} = 1) = E_s(p(x_{ij} = 1|s)) = E_s(p_{ij})$ and $p(x_{ij} = 0) = E_s(q_{ij})$. The information gain provided by $\mathcal{E}_{ij}$ is:

$$\begin{aligned} &I(\mathcal{E}_{ij}, Pr(s)) = E_s[p_{ij} log(\tfrac{p_{ij}}{p(x_{ij}=1)}) + q_{ij} log(\tfrac{q_{ij}}{p(x_{ij}=0)})] \\ &= E_s[p_{ij} log(p_{ij}) + q_{ij} log(q_{ij})] \\ &- E_s(p_{ij}) log(E_s(p_{ij})) - E_s(q_{ij}) log(E_s(q_{ij})) \end{aligned} \quad (15)$$

It is worth noting that the prior distribution $Pr(s)$ is actually conditioned on previous observations as in Eq. 12, but we omit the conditions here for ease of representation. In Eq. 12, we introduced a Gaussian approximation to the posterior distribution. Therefore, we can use the Gauss-Hermite quadrature [15] to compute the expectation efficiently. Fig. 1 shows how the EIG $I(\mathcal{E}_{ij}, Pr(s))$ changes with the expectation and the standard deviation of $s_i - s_j$. We can see that the utility of the PC test increases as $E(s_i - s_j)$ decreases and $std(s_i - s_j)$ increases. This implies that the EIG obtained by performing a PC test on two images with similar quality is higher than that for those with very different quality.

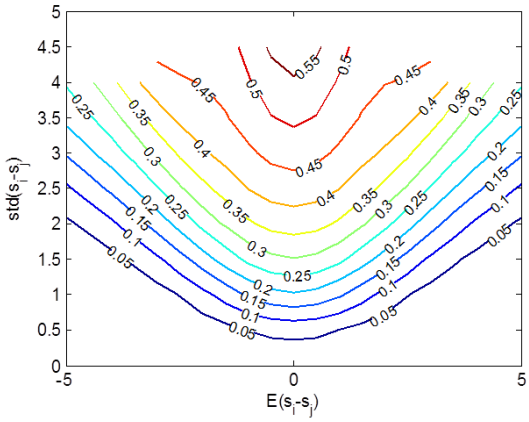

Figure 1: Contour plot of $I(\mathcal{E}_{ij}, Pr(s_i, s_j))$ as function of the expectation and the standard deviation of $s_i - s_j$.

4.2. Active Sampling

Suppose we have n images, we would like to make one observation related to each image at each round of the experiment. For an image A_i, n different tests related to A_i can be conducted including one MOS test $\mathcal{E}_i$ and $n-1$ PC tests $\mathcal{E}_{ij}, j = 1, ...n, j \neq i$. We select the test which has the highest EIG to perform. At the t-th iteration of the experiment, the test selected to perform for image A_i is:

$$\mathcal{E}^t(A_i) = argmax_{\mathcal{E} \in \{\mathcal{E}_{ij}, \mathcal{E}_i | j \neq i\}} I(\mathcal{E}, Pr(s|M_{t-1}, P_{t-1})) \tag{16}$$

where M_{t-1} and P_{t-1} summarize all rating and preference observations in the previous iterations of the experiment.

The proposed model assumes that the observation noise has the same standard deviation $\sigma = 1/\sqrt{2}$ for both the MOS test and the PC test. However, in practice, this assumption may not be true and the MOS test is usually associated with higher noise levels. Therefore the utility of performing a MOS test computed under this assumption may be higher than its true utility. A quick fix can be applied by imposing a slightly higher σ when computing the EIG for the MOS test. In particular, we set $\sigma_{mos} = \frac{1}{\sqrt{2}}(1+\alpha)$, where α is a small positive constant. It is worth noting that α is the only system parameter that is relatively sensitive and needs to be carefully specified in the experiment.

5. Experiments

5.1. Dataset

We have not found any publicly available dataset with a large set of rating and preference judgments for crowdsourcing QoE problems, so we built our own dataset starting with the LIVE IQA dataset [17]. The LIVE IQA dataset includes 779 distorted images with five different types of distortions derived from 29 reference images. For this work, we selected a subset of 120 images from the Fast-Fading category. The 120 images include 20 undistorted reference images and 100 distorted images derived from the 20 reference images. Each image in the LIVE dataset is associated with a subjective DMOS score which was obtained through the MOS test. Note that we will not use the DMOS as groundtruth for our experiments, since the accuracy of the DMOS is limited by the nature of the MOS test. Alternatively, we will generate more realistic groundtruth through the proposed experimental design.

The subjective judgments of the set of images were obtained using the Amazon Mechanical Turk (MTurk) platform[2]. In the MOS test, images are labeled by five ordinal scales: "Bad", "Poor", "Fair", "Good" and "Excellent". For each image, we collected 50 rating observations and a total of 6000 rating scores for 120 images were obtained from 86 subjects. A complete set of paired comparisons of this dataset includes $\binom{120}{2} = 7140$ pairs. For each pair, we collected five repeated observations for a total of 35700 pairs from 196 subjects.

Using MTurk, each working session is considered one HIT (human intelligence task). In our studies, each HIT includes 10 images for the MOS test or 10 pairs for the PC test. Images were randomly permuted to display in each HIT and for the PC test, the display order of a pair of images was also randomized. Additionally, the maximal number of HITs of PC tests that could be done by one worker was limited to 40 so that we would not have a large set of paired comparisons from the same subject.

5.2. Evaluation Measure

We used the Kendall's τ Coefficient and Linear Correlation Coefficient (LCC) for evaluating the performance of subjective tests. Given two global scores on a set of images x_i and y_i $i = 1, ..., n$, Kendall's τ coefficient is defined as

$$\tau(x, y) = \frac{\sum_{i \neq j} X_{ij} Y_{ij}}{\frac{1}{2} n(n-1)} \tag{17}$$

where $X_{ij} = sign(x_i - x_j)$ and $Y_{ij} = sign(y_i - y_j)$. A pair is concordant if $X_{ij} = Y_{ij}$ and is discordant otherwise. $\tau(x, y)$ measures the percentage of concordance pairs minus the percentage of discordant pairs. For identical rankings $\tau(x, x) = 1$ and for reversed rankings $\tau(x, -x) = -1$.

LCC estimates the strength of the linear relationship between x and y. A high value of $LCC(x, y)$ does not necessarily imply a high $\tau(x, y)$. Kendall's τ coefficient is a stricter measure in that it is based on pairwise comparisons.

5.3. GroundTruth

Groundtruth is obtained by solving the MAP problem described in Section 3.3 given all observations, including

[2]https://www.mturk.com/

(a) (b)

Figure 2: Example of Image Pairs.

6000 rating observations and 35700 preference observations. A smoothing constant 0.5 was added to each zero-valued entry in the preference matrix P obtained from the PC test. The observation matrices M and P are normalized so that $P(i,j) + P(j,i) = 1$ and $\sum_{k=1}^{5} M(k,i) = 1$. We use Ipopt [22] for computing the MAP estimate of the underlying score, i.e the minimizer of Eq. 8. The prior distribution of the underlying score is specified by an uninformative prior defined $N(\mu, \Omega)$, where $\mu = \mathbf{0}, \Omega = 1000 \times \mathbf{I}$ and $\mathbf{I}$ is an identity matrix. The correlations between the estimated score obtained from our crowdsourced data and the DMOS data provided in the LIVE dataset are $LCC = 0.978$ and $SROCC = 0.859$. We can see that the LCC value is high, but the Kendall's τ correlation value is relatively low. Examples of image pairs that DMOS disagrees with our estimates are shown in Fig. 2. In this case, the two images are fairly close in quality. In our studies, the first image is preferred to the second image. Taking a closer look at these two images, the first image preserves more detailed information and the second image is more blurred. However, the first image has slightly higher ringing effect compared to the second image. Making a preference judgment between these two images is a highly subjective task and different subjects may have different preferences. In our study, it seems the subjects are more sensitive to blur distortion and prefer sharper images.

5.4. Evaluation

To test the performance of our hybrid system with active sampling, we simulate the crowdsourcing experiment by repeatedly and randomly sampling from real judgments collected from MTurk. All the raw data is included and no post-screening process is applied, because we want to simulate a real crowdsourcing scenario where we do not have enough data to evaluate the rater's reliability for several initial rounds of the experiment .

At each round of the experiment 120 observations were obtained using four different methods:

HY-ACT: The proposed hybrid system with active sampling (described in Section 4) and $\alpha = 0.2$.

HY-RND: A hybrid system using the random sampling method. For image A_i, with probability 0.5 that the MOS test $\mathcal{E}_i$ is sampled, and with probability 0.5 a PC test is sampled and it is uniformly randomly chosen from $\{\mathcal{E}_{ij} | j = 1, .., n, j \neq i\}$;

MOS: A standard MOS test, where at each iteration of the experiment, we make one additional MOS observation for each image.

HRRG [24]: A standard PC test with random sampling. At each iteration, 120 random pairs are sampled based on Erdös-Rényi random graphs.

After each iteration of the experiment, estimates of the underlying scores are obtained using all previously observed data. In particular, HY-ACT and HY-RND estimate the underlying scores by solving the MAP problem described in Section 3.3. MOS simply takes the average of all observations for one particular image as its score. HRRG uses the HodgeRank [11] with an angular transform model to obtain the underlying score. In the first iteration of the experiment, HY-ACT and HY-RND were initialized with 120 MOS tests. After initialization, 150 rounds of experiments were preformed and in the end a total of $151 \times 120 = 18120$ observations were obtained. In this experiment, we simply set $\gamma = \{1, 2, 3\}$ since we found that with this approximation the performance of the proposed method does not vary much and it is faster to run the experiment. The process was repeated 100 times and the median values of the Kendall's τ correlation and LCC are presented in Figs. 3 and 4, where the x-axis represents the number of observations for all 120 images. When the number of observations is very small (for example, less than 10 observations can be made for each image), the HY-ACT curve and the MOS curve are almost identical. This is because at the first several rounds of the experiment, MOS tests have higher EIGs than PC tests and

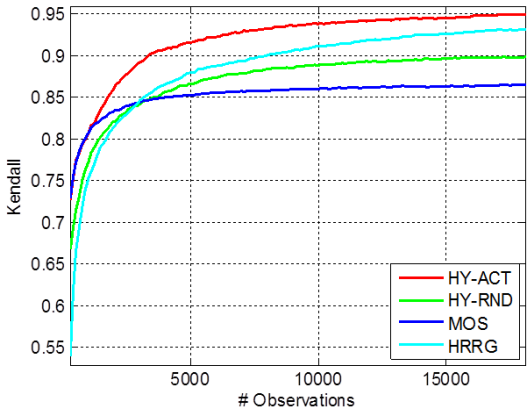

Figure 3: Kendall's τ in the Crowdsourcing experiment.

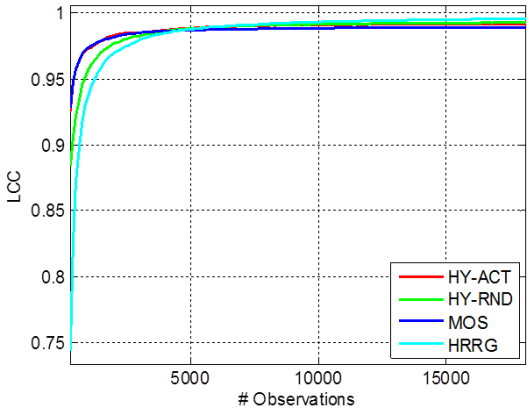

Figure 4: LCC in the Crowdsourcing experiment.

all 120 selected tests are MOS tests. As more observations are obtained, the active sampling method starts to sample more PC tests. Fig. 5 shows the average number of MOS and PC tests performed at each iteration of the experiment.

Due to high observation noise associated with the MOS test, the Kendall's τ coefficients of the MOS test is low even with a large number of rating observations and the PC test is indeed more accurate than the MOS test. The active sampling is critical to the success of the hybrid test, since when a random sampling method as in HY-RND is used, the performance of the hybrid system drops. Table 1 shows the average number of observations required for each image to achieve a given Kendall's τ coefficient. Compared to HRRG, HY-ACT significantly reduced the required number of observations to achieve a given accuracy. Fig. 6 shows the standard deviation (STD) of the Kendall's τ coefficients of the 100 repeated experiments. We can see that HY-ACT has smaller STD than other methods, which implies that HY-ACT has more consistent performance and is thus more reliable.

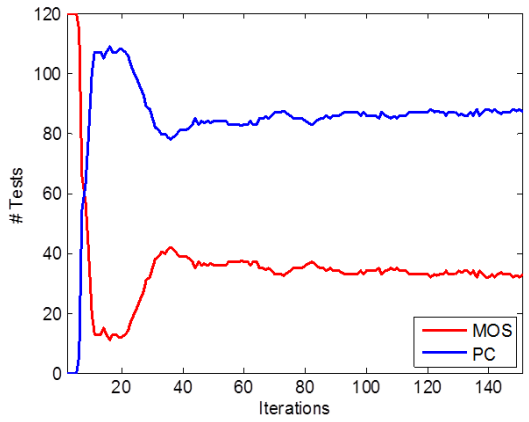

Figure 5: Number of MOS and PC tests sampled in each iteration.

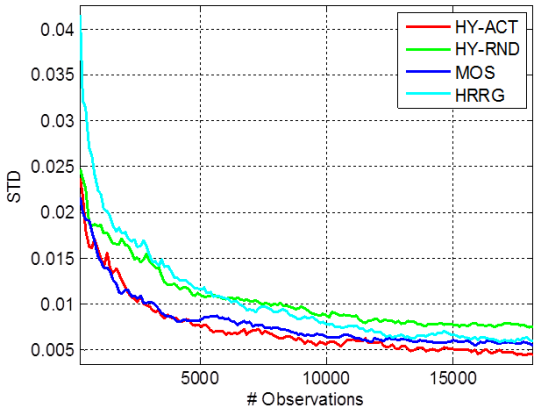

Figure 6: Standard deviation of Kendall's τ.

Kendall	0.85	0.90	0.91	0.92	0.93
HRRG	27	67	82	107	138
HY-ACT	15	28	36	47	61

Table 1: Average number of required observations per image for achieving a given Kedall's τ.

6. Discussions and Conclusions

The proposed model assumes that the variances of observation noise for different images in different types of tests are the same. This assumption does not necessarily hold in practice. Our future work will extend this model to take into consideration these factors. In our experiments, we have used an uninformative prior for the underlying score, however, when additional information about the underlying score is available, it can easily be incorporated into our model by constructing the prior distribution using prior information. The current model can also be extended to an online learning setting using the techniques introduced in

[21].

We have presented a hybrid system which combines the MOS test and the PC test via a unified probabilistic model for estimating the underlying quality scores of images. An active sampling method has been introduced to efficiently construct queries of tests which maximize the expected information gain. The proposed method effectively reduced the required number of observations for achieving a certain accuracy and improved on the state of the art.

Acknowledgment

The partial support of this research by DARPA through BBN/DARPA Award HR0011-08-C-0004 under subcontract 9500009235, the US Government through NSF Awards IIS-0812111 and IIS-1262122 is gratefully acknowledged.

References

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

[2] A. Azevedo-Filho and R. D. Shachter. Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In *Uncertainty in Artificial Intelligence*, pages 28–36, 1994.

[3] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[4] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: preference judgments for relevance. In *Proceedings of the IR research, 30th European Conference on Advances in information retrieval*, pages 16–27, Berlin, Heidelberg, 2008.

[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*, pages 491–500, 2009.

[6] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 193–202. ACM, 2013.

[7] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, Dec. 2005.

[8] H. David. *The Method of Paired Comparisons*. Hodder Arnold, second edition, 1988.

[9] J. Ford, L. R. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8):pp. 28–33, 1957.

[10] D. F. Gleich and L.-h. Lim. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 60–68, 2011.

[11] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

[12] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):pp. 986–1005, 1956.

[13] F. Mosteller. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16:3–9, 1951.

[14] T. Pfeiffer, X. A. Gao, A. Mao, Y. Chen, and D. G. Rand. Adaptive polling and information aggregation. In *The 26th Conference on Artificial Intelligence (AAAI'12)*, 2012.

[15] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C : The Art of Scientific Computing*. Cambridge University Press, October 1992.

[16] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer. Crowdmos: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2416–2419, May 2011.

[17] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.

[18] L. Thurstone. A law of comparative judgment. *Psychological Review*, 1927. 34:273-286.

[19] W. Torgerson. *Theory and methods of scaling*. John Wiley & Sons, New York, 1958.

[20] M. N. Volkovs and R. S. Zemel. A flexible generative model for preference aggregation. In *Proceedings of the 21st International Conference on World Wide Web*, pages 479–488, 2012.

[21] R. C. Weng and C.-J. Lin. A bayesian approximation method for online ranking. *Journal of Machine Learning Research*, 12:267–300, 2011.

[22] A. Wchter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106:25–57, 2006.

[23] Q. Xu, Q. Huang, and Y. Yao. Online crowdsourcing subjective image quality assessment. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 359–368. ACM, 2012.

[24] Q. Xu, T. Jiang, Y. Yao, Q. Huang, B. Yan, and W. Lin. Random partial paired comparison for subjective video quality assessment via hodgerank. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 393–402. ACM, 2011.