This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2831639, IEEE Transactions on Multimedia

Active Learning for Crowdsourced QoE Modeling

Haw-Shiuan Chang, Chih-Fan Hsu, Tobias Hoßfeld, and Kuan-Ta Chen

***Abstract*—Quality of experience (QoE) models predict the subjective quality of multimedia based on the relevant quality of service (QoS) factors. Due to the large space of QoS factors and the high costs of conducting subjective tests, efficient sampling strategies are required to determine which QoS configurations to be queried, i.e., evaluated by subjects.**

In this study, we extend the IQX model proposed by Fiedler et al. [1] towards a multidimensional QoS–QoE model (MIQX). To explore the complicated interaction between QoS factors more efficiently, we develop active learning algorithms for the multidimensional QoE model. Then, we conduct comprehensive experiments to compare the effectiveness of applying different sampling methods to crowdsourced video quality assessment tasks. In offline experiments that assume annotators give the same scores after changing the querying order, we demonstrate that active learning performs best and that a space-filling algorithm performs significantly better than random sampling.

However, when we analyze the performance of the active sampling approaches more deeply using a novel field experiment, we observe that the active learning algorithms, which have been shown to be effective in the offline setting, can fail due to the habituation effect and individual differences of annotators. The active learning methods can also succeed when these issues are mitigated. These findings suggest that simply simulating the sample acquisition order, which is widely adopted in previous active learning literature [2, 3, 4, 5], is not sufficient for multimedia quality assessment tasks.

***Index Terms*—active learning, multidimensional QoS-QoE model, IQX and MIQX model, maximin sampling, crowdsourcing, video quality assessment**

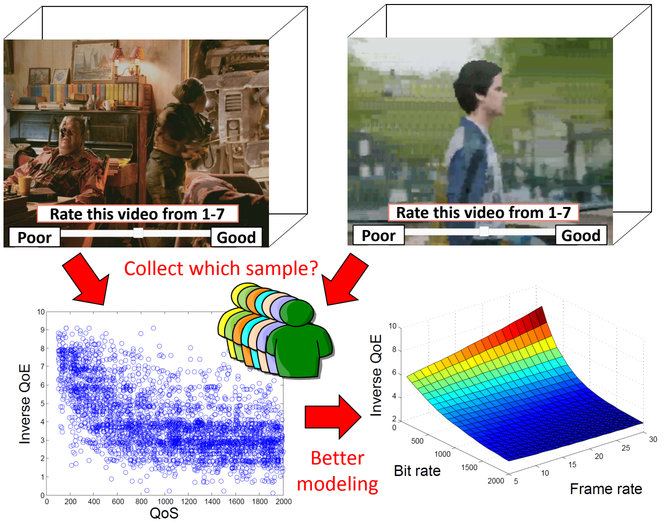


Fig. 1: It is costly to estimate QoE given several QoS factors (e.g., how subjective quality changes as the frame rate increases when using a low bit rate). To model the exponential interdependency between multiple QoS and QoE more efficiently, we update the proposed MIQX model upon receiving a new crowdsourced QoE annotation and select the next query sample that is most informative to the current QoE model via active learning.

I. Introduction

Transmitting multimedia requires substantial Internet bandwidth. As increasing multimedia content becomes available, the optimization of users' experiences on multimedia given limited resources is more important than ever. The problem is challenging in several aspects. First, the mechanisms by which humans judge the quality of multimedia are not fully understood, so the expensive process of collecting subjects' opinions is usually required for satisfactory quality of experience (QoE) estimation. Another major challenge is the many dynamic quality of service (QoS) factors that affect QoE in users' minds [6]. For example, when streaming videos, we often adjust the bit rate or resolution of the videos in response to changing bandwidth or packet delay on the Internet [7].

In QoE modeling, a standard approach is to perform random (or grid) sampling in a large QoS factor space [8], ask annotators to score the quality of those samples, and model QoE–QoS relations using supervised learning [6, 9, 10]. To reduce the cost of collecting annotations, the goal of this article is to actively select samples to better model the relationships between QoS parameters and QoE with fewer samples, as shown in Figure 1.

In previous studies, the relationships between one-dimensional QoS and QoE have been modeled in different ways. Among these models, the IQX hypothesis (exponential interdependency of QoE and QoS) has been shown to be useful in fitting various QoS–QoE relationships [1]. However, the IQX model covers only the case where QoS is one-dimensional. In the real world, strong interactions among multiple QoS factors are common [6, 11, 12], so we extend the IQX model to capture such interactions and call it the multidimensional IQX (MIQX) model.

Due to the potentially huge parameter space of multidimensional methods, efficient approaches are needed to reduce the great expense of measuring multimedia QoE. As verified in [10, 13, 14, 15], crowdsourcing is a cost-effective method to build QoE models. To further reduce the costs of human annotations required in the crowdsourcing setting, we apply different sampling strategies to accurately model QoS–QoE relationships with fewer samples. Furthermore, we propose novel active sampling algorithms based on a probabilistic interpretation of our MIQX model.

In our experiment, we first verify that the MIQX model fits the human judgments well compared with other regression models. Next, after evaluating multiple sampling methods, we

H.-S. Chang is with University of Massachusetts, Amherst, USA (hschang@cs.umass.edu)

Tobias Hoßfeld is with Chair of Communication Networks, University of Würzburg, Germany (tobias.hossfeld@uni-wuerzburg.de)

C.-F. Hsu and K.-T. Chen are with Institute of Information Science, Academia Sinica, Taipei, Taiwan ({chihfan, swc}@iis.sinica.edu.tw)

find that space-filling sampling is nearly always significantly better than random sampling. In addition, we find that our proposed active sampling can outperform space-filling sampling when we assume that the subjects' judgments remain the same when the annotation order changes. We also propose a novel field experiment to assess active sampling methods for QoE estimation, which removes the assumptions. However, in our field experiment, the performance of active learning drops substantially because of the habituation effect and individual differences among subjects. Finally, we integrate space-filling sampling and active sampling and demonstrate that active learning can be employed in crowdsourcing QoE modeling problems.

A. Contributions

- We generalize the IQX model proposed by Fiedler et al. [1] and show that the proposed MIQX model is effective in modeling the relationships between multidimensional QoS and QoE. The probabilistic interpretation of the MIQX model is also introduced.
- We apply active learning to the QoS–QoE modeling problem. The active learning approaches select the next query that minimizes the uncertainty of the QoE model. The experimental results indicate that such a sampling strategy can train much more accurate MIQX models while using the same number of samples in a crowdsourced video quality assessment (VQA) task.
- We propose a field experiment to evaluate different sampling methods for the QoS–QoE modeling problem. Our VQA experiments indicate that space-filling sampling consistently performs better than random sampling, that the performance of active sampling can drop dramatically due to the habituation effect and individual differences among annotators, and that active sampling can be implemented to improve the sampling efficiency when the testing setup is chosen properly (e.g., using double stimulus).

B. Structure of the Article

The remainder of this work is structured as follows. Section II reviews the literature on QoS–QoE modeling with a particular focus on multidimensional QoE models. The state-of-the-art of active learning methods applied to (crowdsourced) QoE studies is revisited. The extended MIQX model is derived and presented in Section III before the adaptive sampling approach is introduced in Section IV. The issues with grid and random sampling are revealed, online space-filling sampling is discussed and several active sampling approaches are introduced. The performance of these approaches is evaluated in Section V based on a video QoE study, which is analyzed offline with respect to model accuracy. Section VI then applies the active sampling method in an online field experiment in which single-stimulus and double-stimulus test designs are investigated. Finally, the limitations of the study and remaining future work are briefly summarized in Section VII. Section IX concludes the work.

II. Related Works

A. QoS–QoE Modeling

The prediction of the subjective quality (QoE) of multimedia based on objective measurements (QoS) is a difficult but important problem. Therefore, abundant studies have attempted to address this challenge. Machine learning methods have recently received increasing attention in QoE modeling and are surveyed by Aroussi and Mellouk [16]. These models provide an output for any given input, and the underlying mathematical model and structure are considered to be a black box. By contrast, white box models provide a mathematical model and a clear relationship between the input and output. Alreshoodi and Woods [17] survey the fundamental relationships between QoS and QoE. Schatz et al. [18] summarizes the available QoE models, focusing on concrete applications including voice communication services, audio-visual services, and web browsing. A recent study by Tsolkas et al. [19] considers parametric QoE estimation for those services and additionally takes into account Skype, IPTV and file download services. Chikkerur et al. [20] addresses media-layer video quality models that also consider speech or video signals in the QoE model, which makes media-layer models much more complex than the parametric models suited for QoE management [18].

Although the relationships between QoE and QoS are intensively studied in the literature, the major focus is on one-dimensional QoS. Hoßfeld et al. [12] investigates multidimensional QoS–QoE models in general and questions whether the multidimensional QoE models for several parameters are additive or multiplicative combinations of the underlying one-dimensional QoE models. The authors show that subjective tests are required to test which type of multifactor QoE model is appropriate for a specific multimedia application and service scenario. However, the generic relationship, and how to best model it, is not yet clear. Due to the growth of the parameter space, where each parameter adds one dimension of complexity, in practice, an appropriate sampling approach and parameter selection method are required.

Our goal is to model the relationships between multidimensional QoS and QoE. The task has been formulated as a curve fitting problem [6], a regression problem [9], and a classification problem [10]. As shown in [1], exponential relationships between QoE and QoS are often observed in practice for various multimedia services, as expressed by the IQX model. The IQX model is promising when a factor dominates QoE. For example, in HTTP video streaming [15, 21], video freezes, i.e., video stalling, dominates user perception and severely degrades QoE. Exponential decays of QoE depending on the number of stalls are observed in subjective studies [15]. On the other hand, previous approaches [6, 9, 10] view the QoS–QoE modeling problem as a general statistical variable mapping problem without considering the prior knowledge of the exponential relationship. Moreover, these approaches prefer more complex discriminative models to achieve better accuracy, but the interpretation and analysis of a complex model is often challenging. The discriminative characteristic also makes active sampling substantially more difficult.

B. Reducing Costs for Subjective QoE Studies

Crowdsourcing enables efficient collection of subjective user ratings in QoE studies transferred from a traditional labo-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2831639, IEEE Transactions on Multimedia

ratory environment onto the Internet. Crowdsourcing has been applied to various multimedia services, e.g., to derive video QoE [13, 14, 15]. The large global pool of subjects results in reduced turnaround times and lower compensation costs for test subjects, and tests can be conducted in the real-world environments of the subjects. However, due to the remote conduction of the experiments, the QoE tests must be designed properly. For example, Hoßfeld and Keimel [22] and Gardlo et al. [23] show that the integration of subject training and appropriate reliability questions in the task design increases the data quality. Best practices and practical guidelines for crowdsourced QoE tests provided by Hoßfeld et al. [13] are also taken into account in our studies in Section VI and Section V.

For crowdsourced QoE tests, only preliminary work, which dynamically selects tests conditions according to the current QoE ratings by the users, exists. Seufert et al. [24] provides an approach to conduct QoE studies that are constrained by a fixed budget for user ratings, i.e., costs for the QoE campaign. Thereby, the next test condition is dynamically selected based on a certain test statistic, which is the confidence interval width of the observed mean opinion scores (MOS) for that test condition. This *adaptive crowdsourcing* approach allocates more user ratings to test conditions with high uncertainty.

Active learning is a machine learning technique that utilizes information to increase the sampling efficiency for model training. The sampling strategy has been shown to be effective in various applications, such as natural language processing, computer vision, and robotics [25]. Active sampling is used by Osting et al. [5] for paired comparison tests to evaluate the user's preference for a pair of video stimuli. For image quality assessment, Ye and Doermann [4] also apply active learning. They propose a method that constructs a set of queries for single-stimulus and paired comparison tests based on the expected information gain provided by each test.

Menkovski and Liotta [3] develops an active learning approach based on maximum likelihood difference scaling (MLDS) to derive a QoE model for video bit rates. The idea of MLDS is to scale differences between test stimuli, which are videos with specific bit rates. Hence, the paired comparison is the QoE test design, and active learning adaptively determines the next pair to be tested.

However, our problem is different. Our goal is to minimize the number of samples required to model the relationships between several QoS parameters and QoE following the MIQX model. The literature aims to minimize the number of samples required to determine the quality of each piece of content (e.g., an image or a video clip) [4, 5] or uses different test designs to model relative differences in quality [2, 3]. Furthermore, the methods are tested using a simulated sampling order, whereas our field experiment demonstrates the limitations of such an evaluation method.

TABLE I: Important notation.

Symbol	Meaning
$\mathbf{x}$	A specific configuration of QoS factors
y	A QoE score labeled by an annotator
$\mathbf{w}$	The learned parameters for all features
α	The largest difference in prediction scores in the QoE model
γ	The minimum prediction score in the QoE model
θ	An abbreviation of all parameters in the QoE model (i.e., $[\alpha\ \gamma\ \mathbf{w}]$)
$\phi()$	The selected kernel function that expands the QoS factors
$f_\theta(\mathbf{x})$	The predicted QoE score of $\mathbf{x}$ using θ as the model parameters
$\sigma^2(f_\theta(\mathbf{x}))$	The variance of the predicted scores given an uncertain θ
$P(\mathbf{x})$	The probability of observing $\mathbf{x}$ in the testing corpus
$\mathbf{y}$	All annotations in the training dataset
$\mathbf{X}$	All configurations of QoS factors in the training dataset
$e^{\Phi(\mathbf{X},\mathbf{w})}$	$e^{\phi(\mathbf{x})\mathbf{w}}$ for all $\mathbf{x}$ in the training dataset
$F_\theta(\mathbf{X})$	$f_\theta(\mathbf{x})$ for all $\mathbf{x}$ in the training dataset
λ	The penalty weight of the regularization term

III. Multidimensional IQX (MIQX) Modeling

When transmitting multimedia, QoS parameters can usually be measured automatically. However, it is difficult to observe the QoE from the viewpoints of users, so our goal is to predict the QoE based on QoS. This task can be formulated as a regression task [9]. Given the QoS parameters of a video as a feature vector $\mathbf{x}$, we can predict the most likely continuous QoE score y. Formally speaking, we want to find a function (i.e., model) f such that

$$y = f(\mathbf{x}), \tag{1}$$

where $\mathbf{x} = [x_j]_{j=1 \dots n}$ is the feature vector, and x_j is the j^{th} QoS parameter. Note that $\mathbf{x}$ should be contained within a valid feature space χ because each QoS parameter should have an upper bound and a lower bound. Before introducing more complicated formulas, we summarize the important notation in Table I.

Fiedler et al. [1] propose the IQX hypothesis, which utilizes exponential functions to approximate f, and show that the function models the relationships between one QoS parameter and QoE well. Specifically, the IQX model formulates f as

$$f(x_1) = \alpha \cdot e^{-\beta \cdot x_1} + \gamma, \tag{2}$$

where α, β and γ are parameters to be estimated.

Intuitively, the exponential function is used because humans often cannot distinguish quality differences when QoS is sufficiently high, which makes the QoE score saturate at a fixed value as QoS increases. For example, one experiment in [1] shows that the IQX model fits the relationships between the Internet bandwidth and the cancellation rate of browsing a website (i.e., the inverse of QoE). Battisti et al. [11], Korhonen et al. [26] also find that the exponential function is a good approximation of f for modeling the impact of video stalls on QoE in HTTP video streaming [15]. The IQX postulates an exponential relationship between QoE and QoS parameters that are directly perceived by the end user, e.g., application-level QoS such as stalling in the case of video streaming QoE directly perceived by video users [15] and network-level QoS such as packet loss in the case of VoIP, leading to speech degradations [1]. Clearly, several network parameters, including network bandwidth and packet loss, and application parameters, such as the video bit rate and the video buffer size, affect stalling. A closed formula $g(\mathbf{x})$ for quantifying the number of stalls and stall duration depending on these parameters [27] may be applied in the IQX equation, $f(g(\mathbf{x})) = \alpha \cdot e^{-g(\mathbf{x})} + \gamma$.

To incorporate more features, we propose the multidimensional IQX model (MIQX) by generalizing the IQX model. Given multiple QoS parameters $\mathbf{x}$, we first use a kernel function ϕ to explore the interactions among features by mapping them into a space with higher dimensions L. Then, the weights $\mathbf{w}$ linearly combine these mapped features into a one-dimensional value $\phi(\mathbf{x})\mathbf{w}$. Finally, an exponential function confines the value to the range of QoE score. The equation can be written as

$$f_\theta(\mathbf{x}) = \alpha \cdot e^{-\phi(\mathbf{x})\mathbf{w}} + \gamma, \tag{3}$$

where $\phi(\mathbf{x})$ and $\mathbf{w}$ are vectors of size $1 \times L$ and $L \times 1$, respectively, and $\theta = [\alpha \ \gamma \ \mathbf{w}]$ are the parameters of the model. Note that the model is an extension of the IQX model (2), where we replace the weighted one-dimensional QoS parameter $\beta \cdot x_1$ with a weighted combination of multiple parameters $\phi(\mathbf{x})\mathbf{w}$.

By viewing the task as a regression problem, we can estimate the parameters in our model, $\theta = [\alpha \ \gamma \ \mathbf{w}]$, via training the model on existing QoS–QoE pairs. In the training data, we denote all N collected labels as an $N \times 1$ vector $\mathbf{y} = [y_i]_{i=1\ldots N}$, where y_i is the i^{th} QoE score, and denote all estimated QoE scores using parameters θ as $F_\theta(\mathbf{X}) = [f_\theta(\mathbf{x_i})]_{i=1\ldots N}$, where $\mathbf{x_i}$ is the QoS parameters for the i^{th} sample.

When training the model, one goal is to minimize the 2-norm errors between the i^{th} estimated QoE score $f_\theta(\mathbf{x_i})$ using parameters θ and the i^{th} observed QoE score y_i for all i. Therefore, we define loss function $E(\theta, \mathbf{y}, \mathbf{X})$ as

$$\begin{aligned} E(\theta, \mathbf{y}, \mathbf{X}) &= \sum_{i=1}^{N} \left(f_\theta(\mathbf{x_i}) - y_i\right)^2 = \left(\mathbf{y} - F_\theta(\mathbf{X})\right)^T \left(\mathbf{y} - F_\theta(\mathbf{X})\right) \\ &= \left(\alpha \cdot e^{-\Phi(\mathbf{X},\mathbf{w})} + \gamma \cdot \mathbb{1} - \mathbf{y}\right)^T \left(\alpha \cdot e^{-\Phi(\mathbf{X},\mathbf{w})} + \gamma \cdot \mathbb{1} - \mathbf{y}\right), \end{aligned} \tag{4}$$

where $\alpha \cdot e^{-\Phi(\mathbf{X},\mathbf{w})} = \left[\alpha \cdot e^{-\phi(\mathbf{x_i})\mathbf{w}}\right]_{i=1\ldots N}$, and $\mathbb{1}$ is an $N \times 1$ vector whose elements are all 1.

The QoE scores sometimes contain heavy noise, especially when they are collected using crowdsourcing. Regularization can prevent our model from overfitting the noise in the data and stabilize the parameter estimation process. We do not know the prior distributions of α and γ, but their range can be computed from the range of QoE scores. Thus, we can define the valid set of parameters as

$$\begin{aligned} \Theta = \{[\alpha, \gamma, \mathbf{w}] \mid \mathrm{QoE}_{\mathrm{min}} \le \gamma \le \mathrm{QoE}_{\mathrm{max}} \wedge \\ 0 \le \alpha \le (\mathrm{QoE}_{\mathrm{max}} - \mathrm{QoE}_{\mathrm{min}})\}, \end{aligned} \tag{5}$$

where $\mathrm{QoE}_{\mathrm{max}}$ and $\mathrm{QoE}_{\mathrm{min}}$ are the maximum and minimum QoE scores subjects can annotate. For $\mathbf{w}$, larger values of its elements usually result in a more unstable model because a small value change in a QoS parameter causes a large prediction difference. By combining our loss function and regularization, we estimate the parameters $\theta^* = [\alpha^* \ \gamma^* \ \mathbf{w}^*]$ that minimize the loss function $E(\theta, \mathbf{y}, \mathbf{X})$ and regularization penalty for $\mathbf{w}$ as

$$\theta^* = \arg\min_{\theta \in \Theta} \sum_{i=1}^{N} \left(f_\theta(\mathbf{x_i}) - y_i\right)^2 + \lambda \mathbf{w}^T \mathbf{w}, \tag{6}$$

where λ is a constant hyperparameter that controls the strength of regularization.

The objective function in (6) is non-linear but differentiable, so we use the trust-region-reflective algorithm to perform optimization [28]. Whenever we collect a new sample, we update our parameter estimation based on (6). The optimization procedure requires an initial estimation point, so we provide our last parameter estimation as the initial point when receiving a new sample.

IV. Adaptive Sampling for QoE Modeling

The collection of human opinions is expensive, so our next problem is how to design the sample presentation order such that we can reduce the number of samples required to build an accurate model. Grid and random sampling are standard approaches for the QoE modeling problem [8] and are widely adopted by subjective quality assessment methods [29]. Since space-filling sampling and active sampling have been shown to be effective in many different applications [25, 30], we apply several of these sampling algorithms to our problem.

A. Issues with Grid and Random Sampling

We usually want our sample distribution to be approximately uniform in the feature space. One popular choice is uniform grid sampling, which partitions the feature space into hypercubes of equal size. This solution is easy to implement, but it requires users to set the number of samples in advance and the number cannot be arbitrary.

When we collect annotations via crowdsourcing, some subjects might not finish all the questions in the task; an online sampling method would be more flexible and useful in practice (i.e., we can always determine the next sample based on previous sampling history). A simple choice is random sampling, which randomly and uniformly acquires the next sample.

The feature space in our multidimensional regression problem can be large because QoE can be affected by many factors, including network conditions and the characteristics of multimedia content. Although random sampling according to a uniform distribution approaches uniform sampling in the long run, some large areas in the sampling space may not covered by any sample when the budget is not sufficient for comprehensive sampling.

Furthermore, we may also collect two very similar consecutive samples when using random sampling. In this case, users typically give the same score to both samples, which wastes the annotation for the second video because it provides redundant information for estimating the parameters in our model.

In the crowdsourcing setting, the number of annotations each subject provides is usually small. In a few samples, a subject might see biased sampling results (e.g., the set of samples contains many more high-quality samples than low-quality samples). This bias may affect the subjects' standard of judgment.

B. Online Space-filling Sampling

As shown in [31], uniform sampling works well for estimating an unknown smooth function in a regression problem. In this work, we adopt the sequential design for maximin sampling proposed in [30, 32]. We select the ith sample $\mathbf{x_i}$ as farther from the chosen samples $\mathbf{x}$ as possible. Specifically, the formula for selecting the next sample is

$$\mathbf{x_i} = \arg\max_{\mathbf{x} \in S} \Big(\min_{\mathbf{x_k} \text{ for } k=1,\ldots,(i-1)} \left(d(\mathbf{x}, \mathbf{x_k})\right) \Big), \tag{7}$$

where d is a function that measures the distance between two samples in the feature space; we use Euclidean distance in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2831639, IEEE Transactions on Multimedia

this work. Note that optimization of the objective function in (7) is difficult because the function is not convex (i.e., local minimums exist) [30] when $S = \chi$ contains infinite valid sample points within the range; therefore, our S is constructed with an ample number of points that are randomly sampled in the feature space according to a uniform distribution.

Compared with other space-filling methods, sequential maximin sampling is easy to implement. Additionally, maximin sampling tends to acquire samples near the boundaries of valid range initially. The behavior is similar to that of active sampling and helpful to constructing a better model with a few samples.

C. Active Sampling

Intuitively, the collected annotations impose constraints on the parameter estimation space. In many applications, active learning methods select the next sample that is most informative for estimating the model parameters [25]. To measure the amount of information each sample provides, we first define the probability distribution of the parameters in the model; then, we can apply active learning algorithms to the probabilistic model.

1) Probabilistic MIQX model: We assume that the distribution of error between the predicted QoE scores and the QoE scores provided by subjects is normal $\mathcal{N}(0, \sigma_\nu)$, which is supported by previous work, such as [8]. We further assume that the prior distributions of α and γ are uniform within their range and that the prior distribution of $\mathbf{w}$ is Gaussian with a mean and covariance matrix of 0 and $\frac{1}{\lambda}I$, respectively, as assumed in the Bayesian model of linear regression described in [33].

Based on these assumptions, we define the probability of having model parameters θ given annotated QoE scores $\mathbf{y}$ for samples $\mathbf{X}$ as

$$\begin{aligned} P(\theta|\mathbf{y},\mathbf{X}) &\propto P(\mathbf{y}|\theta,\mathbf{X})P(\theta|\mathbf{X}) \\ &\propto e^{-(\mathbf{y}-F_\theta(\mathbf{X}))^T(\mathbf{y}-F_\theta(\mathbf{X}))}e^{-\lambda\mathbf{w}^T\mathbf{w}}. \end{aligned} \tag{8}$$

Note that the objective function in (6) is equal to $-\log(P(\theta|\mathbf{y},\mathbf{X}))$; therefore, our parameter estimation is actually a maximum a posteriori probability estimation in the probabilistic interpretation.

2) Uncertainty sampling: One active learning strategy is to sample the most uncertain point for the current model in the feature space. In the regression problem, the uncertainty can be represented by the variance of the prediction by means of the current QoE–QoS model $\sigma^2(f_\theta(\mathbf{x}))$. Thus, uncertainty sampling selects one sample at a time according to $\arg\max_{\mathbf{x}\in S}\sigma^2(f_\theta(\mathbf{x}))$ to minimize the overall prediction variance.

A simple way to estimate $\sigma^2(f_\theta(\mathbf{x}))$ is to split the QoS space into multiple bins and to measure the variance of the annotated scores. The experiments in Seufert et al. [24] demonstrate the promise of this approach when considering a single QoS factor. However, in the case of multiple QoS factors, substantially more annotations are required to obtain reasonable variance estimates, and the size of the bins is difficult to set.

Fortunately, the probabilistic MIQX model is a differentiable model, so the prediction variance at any arbitrary point can be approximated in a closed form. According to [34], the variance of the QoE score prediction can be computed as

$$\sigma^2(f_\theta(\mathbf{x})) = g(\mathbf{x})^T A^{-1} g(\mathbf{x}), \tag{9}$$

where $A = \frac{\partial^2(E(\theta,\mathbf{y},\mathbf{X})+\lambda\mathbf{w}^T\mathbf{w})}{\partial\theta^T\partial\theta}$ is the Fisher information matrix, and $g(\mathbf{x}) = \frac{\partial f_\theta(\mathbf{x})}{\partial\theta} = \begin{bmatrix} \frac{\partial f_\theta(\mathbf{x})}{\partial\alpha} \\ \frac{\partial f_\theta(\mathbf{x})}{\partial\gamma} \\ \frac{\partial f_\theta(\mathbf{x})}{\partial\mathbf{w}} \end{bmatrix} = \begin{bmatrix} e^{-\phi(\mathbf{x})\mathbf{w}} \\ 1 \\ -\alpha\cdot e^{-\phi(\mathbf{x})\mathbf{w}}\phi(\mathbf{x}) \end{bmatrix}$. As derived in [34], after neglecting some higher-order terms, selecting the sample with the maximum $\sigma^2(f_\theta(\mathbf{x}))$ is approximately equivalent to maximizing the total information gain measured by the entropy of the parameter distribution. This approach is also known as the c-optimal design [25].

3) Q-optimal design and MMIG: In our MIQX model, the sample with the highest prediction variance is usually the sample on the edge of the valid feature space. This characteristic is often undesirable because the estimated model might suffer from outliers more easily [33], which is a common problem with uncertainty sampling for regression [25, 33, 34]. Intuitively, this problem occurs because the sampling method attempts to find the maximum information gain for the global prediction model $f_\theta(\mathbf{x})$, which can take all possible inputs $\mathbf{x}$. However, the outputs of the model with invalid inputs (e.g., ones with very large or small values) are not relevant for our purpose.

Therefore, it may be more practical to consider the probability of observing QoS parameters $P(\mathbf{x})$ and to focus on minimizing the uncertainty of the prediction from highly probable $\mathbf{x}$. This idea is the basis of density-weighted methods [25]. In this case, minimizing the prediction variance and maximizing the information gain leads to two different sampling strategies: Q-optimal and mean marginal information gain (MMIG).

The goal of Q-optimal [35] is to minimize the variance weighted by the feature distribution as

$$V(\mathbf{X},\mathbf{y}) = \int_{\mathbf{x_u}} P(\mathbf{x_u})\left(\sigma^2(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x_u}))\right)d\mathbf{x_u}, \tag{10}$$

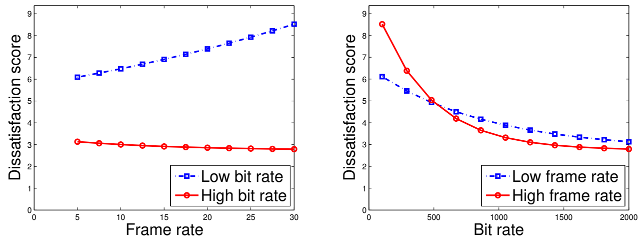
where $\theta(\mathbf{y},\mathbf{X}) = \arg\max_\theta P(\theta|\mathbf{y},\mathbf{X})$ is the model parameters estimated given training data $\mathbf{X}$ and $\mathbf{y}$. Then, we choose the next sample based on the following approximation [34]:

$$\begin{aligned} \mathbf{x_i^*} &= \arg\max_{\mathbf{x_i}\in\mathbf{S}}\left(V(\mathbf{X},\mathbf{y}) - V(([\mathbf{X};\mathbf{x_i}],[\mathbf{y};y_i]))\right) \\ &\approx \arg\max_{\mathbf{x_i}\in\mathbf{S}}\int_{\mathbf{x_u}} P(\mathbf{x_u})\frac{g(\mathbf{x_i})^T A^{-1} g(\mathbf{x_u})}{\sigma^2(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x_i})) + \sigma_\nu^2}d\mathbf{x_u}, \end{aligned} \tag{11}$$

where g, A, and $\sigma^2(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x}))$ are defined in (9), σ_ν^2 is the global variance estimated by $E(\theta,\mathbf{y},\mathbf{X})/N$, and S is the valid set of QoS parameters. Note that we use the same S as in maximin sampling to make the comparison in our experiments fair.

By contrast, MMIG [34] minimizes the uncertainty of the prediction probability distribution $P(f_\theta(\mathbf{x_u}))$. We measure the average uncertainty over all valid features $\mathbf{x_u}$ on the basis of entropy as

$$U(\mathbf{X},\mathbf{y}) = \int_{\mathbf{x_u}} P(\mathbf{x_u})\ \mathrm{ent}(P(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x_u})))d\mathbf{x_u}, \tag{12}$$



(a) The frame rate of videos [1/s] versus the dissatisfactory score given a bit rate of 100 kbps or 2000 kbps.

(b) The bit rate of videos [kbps] versus the dissatisfactory score given a frame rate of 5 fps or 30 fps.

Fig. 2: The prediction results on grid points demonstrate the strong interaction between frame rates and bit rates. The model is estimated using the random samples collected in section V-B. The visualized MIQX model considers only frame rates, bit rates, and their interaction term for predicting video quality.

where $\text{ent}(P(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x_u})))$ is the entropy of the parameter distribution, which can be computed as $\frac{1}{2}\log\left(2\pi e\sigma^2(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x_u}))\right)$ because we assume that our prediction distribution $P(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x_u}))$ is normal. Then, we can determine the next query by maximizing the MMIG as

$$\mathbf{x_i^*} = \arg\max_{\mathbf{x_i}\in\mathbf{S}} (U(\mathbf{X},\mathbf{y}) - U([\mathbf{X};\mathbf{x_i}],[\mathbf{y};y_i])) = \arg\max_{\mathbf{x_i}\in\mathbf{S}} (-\frac{1}{2}\int_{\mathbf{x_u}} P(\mathbf{x_u})\log\left(1-\frac{g(\mathbf{x_i})^T A^{-1} g(\mathbf{x_u})}{\sigma^2(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x_u}))(\sigma^2(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x_i}))+\sigma_\nu^2)}\right)d\mathbf{x_u}), \tag{13}$$

where the notations have the same meaning as in (11).

The probability of observing QoS parameters $P(\mathbf{x_u})$ is highly dependent on the context of the application. The setting of $P(\mathbf{x_u})$ allows us to focus on specific regions or sample points when building the model. For example, if we want to minimize the prediction variance on some grid points within a specific range, we can set $P(\mathbf{x_u}) = 0$ when $\mathbf{x_u}$ is not on those grid points.

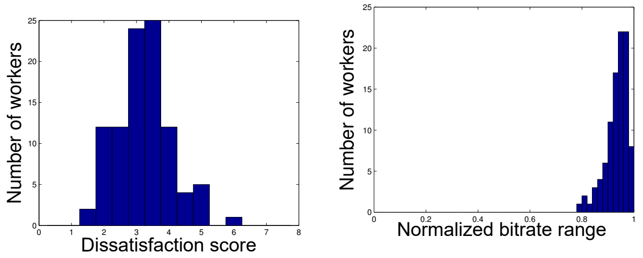
For simplicity, in this work, we assume that $P(\mathbf{x_u})$ is a uniform distribution over the whole valid feature space χ and that the integrals in (11) and (13) are approximated by summation over several thousands of uniformly sampled points in the feature space.

V. Performance Evaluation

To investigate the effect of various sampling methods on different models, we first collect QoE scores from video clips with randomly selected QoS parameters. Then, the sampling methods are evaluated by changing the collection order of subjective annotations offline, as in previous work [2, 3, 4, 5].

A. Experimental Design

Like the experiments in Mushtaq et al. [10], we use the video characteristics as our QoS parameters and perform the QoE in video quality assessment (VQA). The characteristics we consider include bit rate, frame rate, resolution, temporal complexity of videos, and spatial complexity of a video segment, where the complexities of a segment are computed as the summation of the entropies over all color channels to measure the difficulty of compressing the video segment. Specifically, spatial complexity is the average entropy over all frames, and temporal complexity is the average entropy of the color differences of consecutive frames over the whole segment. All features are normalized into the range [0,1], so our feature space χ is a hypercube.



(a) Distribution of average scores.

(b) Distribution of the normalized bit rate range.

Fig. 3: The average score from each subject and the difference between the highest and lowest normalized bit rate that each subject sees within the 30 rounds of annotation.

In the formulation of the IQX model, we assume that the score decreases as the feature values increase. When estimating video quality, we use the MIQX model to describe the decrease in user dissatisfaction as QoS increases, as suggested by Fiedler et al. [1] in their final experiment. Therefore, we use the inverse of QoE scores as the prediction target of our regression task, and the inverse scores are called the dissatisfaction score.

Significant interactions between features exist in VQA. For example, we visualize the interaction between the bit rate and the frame rate of videos by means of the MIQX model in Figure 2. When the bit rate is sufficiently high, the video quality improves (i.e., the dissatisfactory score is lower) when we increase its frame rate. On the other hand, when the bit rate is low, a low frame rate may reduce the distortion introduced by compression at a low bit rate. We add the 2nd-order interaction terms to the model in our MIQX model (i.e., $\phi(\mathbf{x}) = ([x_i]_{i=1\ldots5}, [x_j \cdot x_k]_{j,k=1\ldots5})$), so the number of features L after the mapping is $5 + \binom{5}{2} = 15$.

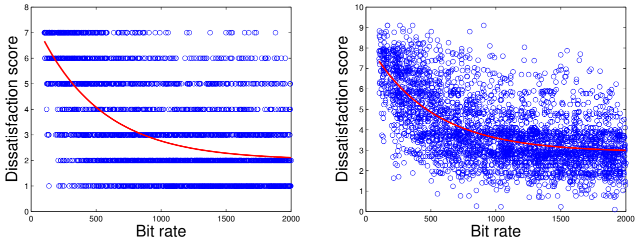
B. Dataset

We collect 3,318 annotations from 97 subjects using Amazon Mechanical Turk (MTurk)[1] and Bounty Worker[2]. We do not consider the network condition in this experiment, so each video is pre-loaded to each subject's computer before being played. A 10 second clip is first shown in full-screen mode; then, subjects are asked to score the quality of the stimulus on a 7-level ordinal scale: "Very Satisfied", "Satisfied", "Somewhat Satisfied", "Neither Satisfied nor Dissatisfied", "Somewhat Dissatisfied", "Dissatisfied", and "Very Dissatisfied". A higher score indicates better quality.

Each clip is randomly chosen from two open-source films: Big Buck Bunny [36] and Tears of Steel [37]. The clip is compressed by H.264 using encoder x264 [38], the bit rate is

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

[2] http://bountyworkers.net/



(a) Before bias removal.

(b) After bias removal.

Fig. 4: The scatter plots depict the dissatisfaction scores of subjects under different bit rates. The red curves are the estimated IQX models with a single feature. The IQX model fits better after removing subject bias (relative squared error (RSE) decreases from 0.64 to 0.59).

randomly sampled from 100 kbps to 2000 kbps, the frame rate is randomly sampled from 5 fps to 30 fps, and the resolution is randomly selected from six different 16:9 resolutions with height 480, 600, 720, 840, 960, and 1080.

The crowdsourcing data suffer from severe subject bias (some subjects tend to score videos higher while some subjects tend to score videos lower). Figure 3a shows the distribution of the average scores from all 97 subjects, and Figure 3b shows the distribution of normalized bit rate ranges. Many average scores deviate substantially from “Neither Satisfied nor Dissatisfied” (i.e., score 4) even though all subjects see nearly the full range of bit rates. To remove the bias and align the scores from different subjects, we standardize the average score of each subject to 4 by shifting all the scores, as suggested by Janowski and Pinson [39]. The normalization also makes our estimated model more robust against cheating outliers (e.g., always thinking the video is satisfied). The effect of this post-processing is visualized in Figure 4.

Each subject has to score at least 30 rounds of videos clips to receive the reward, and we discard the annotations from subjects who do not complete the task. The sufficient number of rounds allows us to measure the scoring bias and to detect cheaters. A subject is more likely to be a cheater if his/her opinions are usually different from the average scores of other subjects. On the basis of this tendency, we apply an approach similar to [40] to filter outliers. However, we cannot find clear boundary between cheaters and normal subjects, and the conclusions of the following experiment results do not change when some of potential cheaters are excluded. The filtering model requires additional hyperparameters, which complicate the analysis, so we report the results based on the annotations from all subjects.

C. Evaluation Method

Given the same query, there is often large variance in crowdsourced scores for VQA, so the performance of the estimated models also varies considerably when the sample size is small. To evaluate the effectiveness of different sampling methods, we conduct 200 trials and make each method collect different samples in each trial by injecting some randomness into the sampling process. At the beginning of a trial, 70% of the randomly selected samples are placed in the training pool, and the rest of the samples are placed in the testing pool. Next, we randomly select 10 samples from the training pool for each sampling method, and let the method choose the next query based on the previously collected samples. During the selection process of a trial, a sample can only be chosen once by a sampling method.

The regression models and sampling methods are evaluated based on the similarity between our predictions and the annotations in the testing pool. We measure the similarity using the relative squared error (RSE), linear correlation coefficient (LCC), and Spearman rank-order correlation coefficient (SROCC). The RSE is defined as $\frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$, where y_i is the ith score rated by subject, $\bar{y}$ is the mean of y_i over all i, and $\hat{y}_i$ is the predictions from different models bounded by the maximum score and the minimum score that subjects can provide.

In addition to the prediction accuracy, we evaluate the accuracy of the parameters in the MIQX model estimated by different sampling methods. The accuracy is measured by the root-mean-square error (RMSE) normalized by the root mean square of the ground truth weights. Specifically, the normalized RMSE is defined as $\frac{\|\hat{\mathbf{w}} - \mathbf{w}\|_2}{\|\mathbf{w}\|_2}$, where $\hat{\mathbf{w}}$ is the inferred weights and $\mathbf{w}$ is the ground truth weights. The ground truth weights refer to the weights estimated based on all the collected samples.

D. Regression Models

First, we verify that our MIQX model is effective in capturing the complex relationships between QoS and QoE in videos. The features and labels are described in Section V-A. Before subject normalization, the range of dissatisfactory scores is from 1 (Very Satisfied) to 7 (Very Dissatisfied). In the following experiment, we constrain α and γ within the ranges [0,12] and [-2,10], respectively, because the range of normalized scores is [-2,10]. The regularization scale λ in the MIQX model is fixed at 0.05.

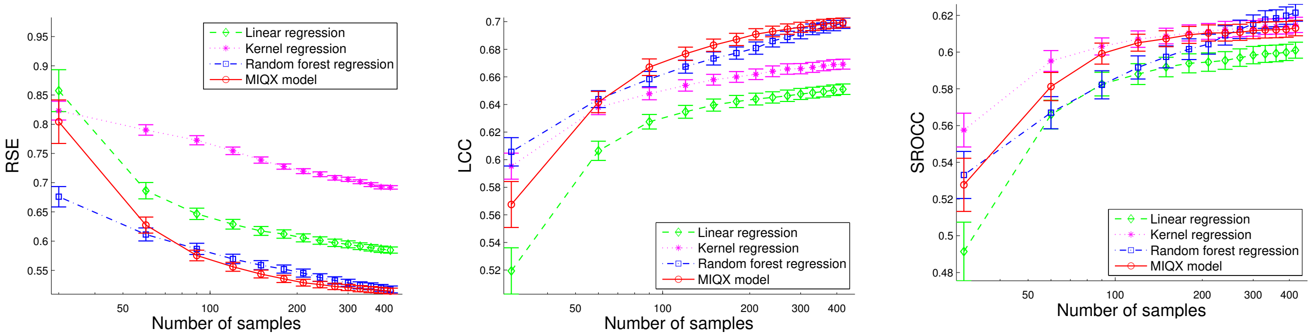
In Figure 5, we compare the MIQX model with widely used regression models, including linear regression with 2nd-order interaction terms, Nadaraya-Watson kernel regression with Gaussian kernel [41], and random forest [42], which is the most accurate QoS–QoE modeling algorithm in Mushtaq et al. [10]. The 2nd-order interaction terms do not increase the performance of the kernel regression and random forest models, so we use only the 5 features for these 2 models.

The performance of the MIQX model is significantly better than that of the simpler models, such as linear regression, whereas it is similar to that of complex models, such as random forest. Therefore, our extension of the IQX hypothesis is a good generative model that can accurately describe the judgments of the crowd in the VQA task.

E. Maximin versus Random Sampling

Two sampling methods are commonly used when performing VQA: random sampling under a uniform distribution and space-filling sampling (e.g., grid sampling). If the sample size is sufficiently large, the sample distributions of both methods should be similar. However, when fewer samples are collected, the prediction accuracies of the methods can be very different.

We investigate the performance difference between maximin sampling (a type of space-filling sampling) and random sampling. In Figure 6, the performances of the simple model and

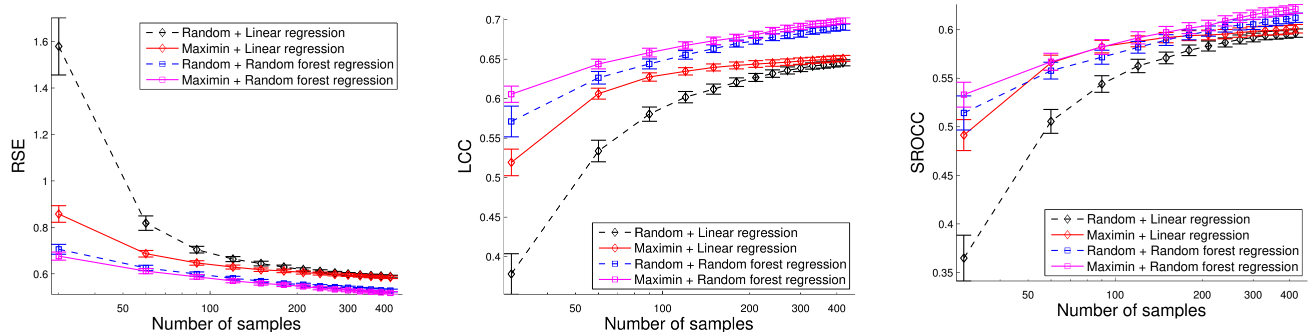


(a) Relative squared error.

(b) Linear correlation coefficient.

(c) Spearman rank-order correlation coefficient.

Fig. 5: The performances of different regression models under maximin sampling. The error bars represent the interval within plus or minus one standard error of the mean. Note that smaller errors and larger correlation coefficients indicate better performance.

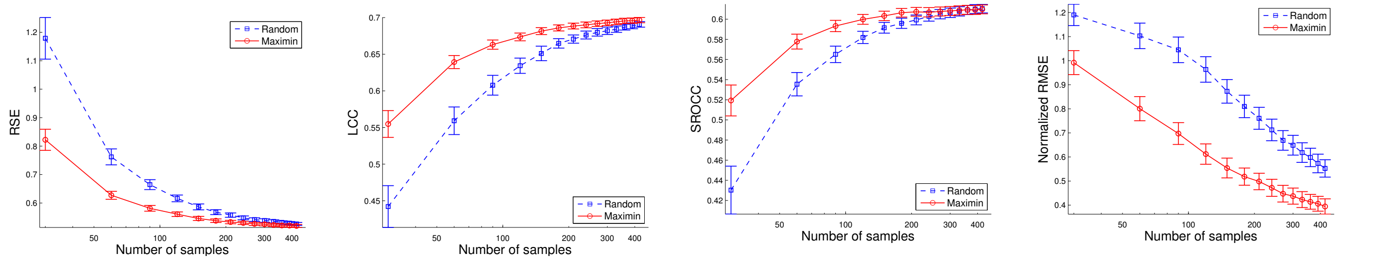


(a) Relative squared error.

(b) Linear correlation coefficient.

(c) Spearman rank-order correlation coefficient.

Fig. 6: Random sampling versus maximin sampling for linear regression and random forest.



(a) Relative squared error.

(b) Linear correlation coefficient.

(c) Spearman rank-order correlation coefficient.

(d) Normalized root-mean-square error of feature weights.

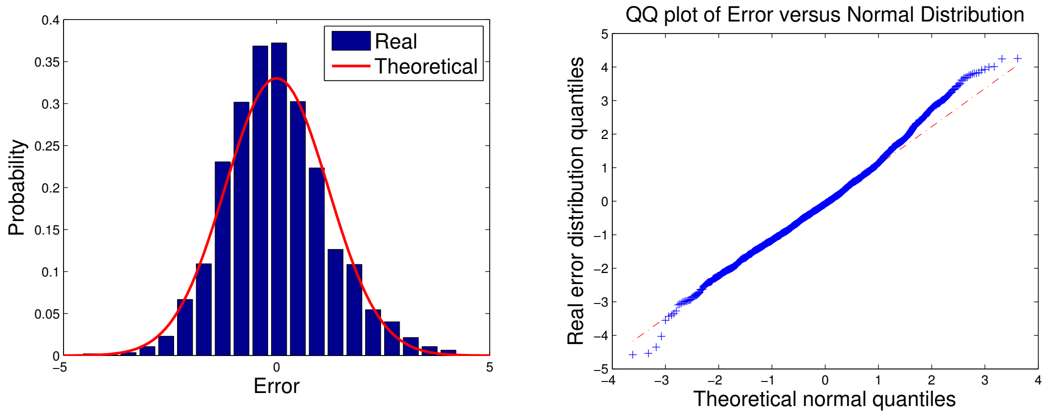
Fig. 7: Random sampling versus maximin sampling for the MIQX model.

complex model (i.e., linear regression and random forest regression) are compared, and we can see that maximin sampling leads to more accurate models. Similarly, the performance of the MIQX model is improved significantly by using maximin sampling compared with that using random sampling, as shown in Figure 7. Note that our estimation of the MIQX model requires initialization. In all sampling methods, the MIQX model is updated upon receiving a new sample, with the previous estimate used for initialization.

F. Active versus Maximin Sampling

As shown in Figure 8, the error distribution can be approximated by a Gaussian distribution, which confirms the assumption made for active sampling algorithms in Section IV-C. Figure 9 shows that active sampling, including Q-optimal and MMIG, outperforms maximin sampling in terms of RSE, LCC, and SROCC. These two methods reach similar normalized RMSE in the parameter space with 200 samples as that achieved by maximin sampling using 400 samples. Note that Q-optimal performs slightly better than MMIG in terms of RSE because the objective function of Q-optimal minimizes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2831639, IEEE Transactions on Multimedia



(a) Histogram showing the empirical error distribution and the red curve showing the theoretical normal distribution.

(b) The two distributions are similar if the points in the QQ plot are close to the straight line.

Fig. 8: Comparison of the error distribution of the MIQX model and the normal distribution.

the squared error of the predictions. The results support the effectiveness of proper active sampling for the MIQX model in an offline setting, which assumes that changing the sampling order does not affect the subjects' scoring judgment.

The uncertainty sampling (c-optimal) does not perform as well as the other two active methods in terms of RSE and normalized RMSE when the number of samples exceeds 60. The reason can be seen in our visualization of the sample locations shown in Figure 10. The c-optimal method samples too many examples with low bit rates because they are usually the locations with higher prediction variance. These sampling locations support our claim in section IV-C3 that uncertainty sampling tends to sample points near the boundary of the feature space. It is worth noting that the samples do not lie exactly on the boundaries because our sampling set S contains only samples collected offline and we do not allow the algorithms to select the same point in each trial.

VI. Field Experiment

A. Experimental Design

In the previous section, we have shown that active sampling methods are useful for estimating the MIQX model in an offline setting. However, we want to verify whether active sampling methods for VQA can successfully reduce the required sample size while achieving the same accuracy in a realistic online setting. To achieve this goal, we create several trials for each sampling method and randomly assign each subject to a trial. In each trial, the query for each subject is determined online based on the previous queries in the same trial. The subject rates 40 samples to complete the task, and incomplete annotations are discarded. Each trial includes one subject at a time until 5 participants complete the task (i.e., 200 samples are collected).

Regardless of the sampling method, the first 10 queries for each subject are randomly selected to align his/her expectation of the video quality. Active sampling must estimate the parameters of the MIQX model based on all previous annotations of subjects in a trial to determine the next query, and the estimation requires the average score of each subject to remove the subject bias online. Therefore, we update the average score of each subject whenever a new sample is collected, and shift all scores of the subject based on the updated average score.

All annotations are crowdsourced from MTurk. To prevent subjects waiting a long time for video compression, all the videos are compressed in advance. Each sampling method selects samples from the pre-compressed clips. We uniformly sample 3,000 QoS parameters as our valid feature set S and create clips using the sampled parameters. Note that the ranges of parameters and videos are the same as those in the offline experiments, except we increase the maximum bit range to 2500 kbps because the quality at a bit rate of 2000 kbps might not be sufficient when the motion in a clip is rapid.

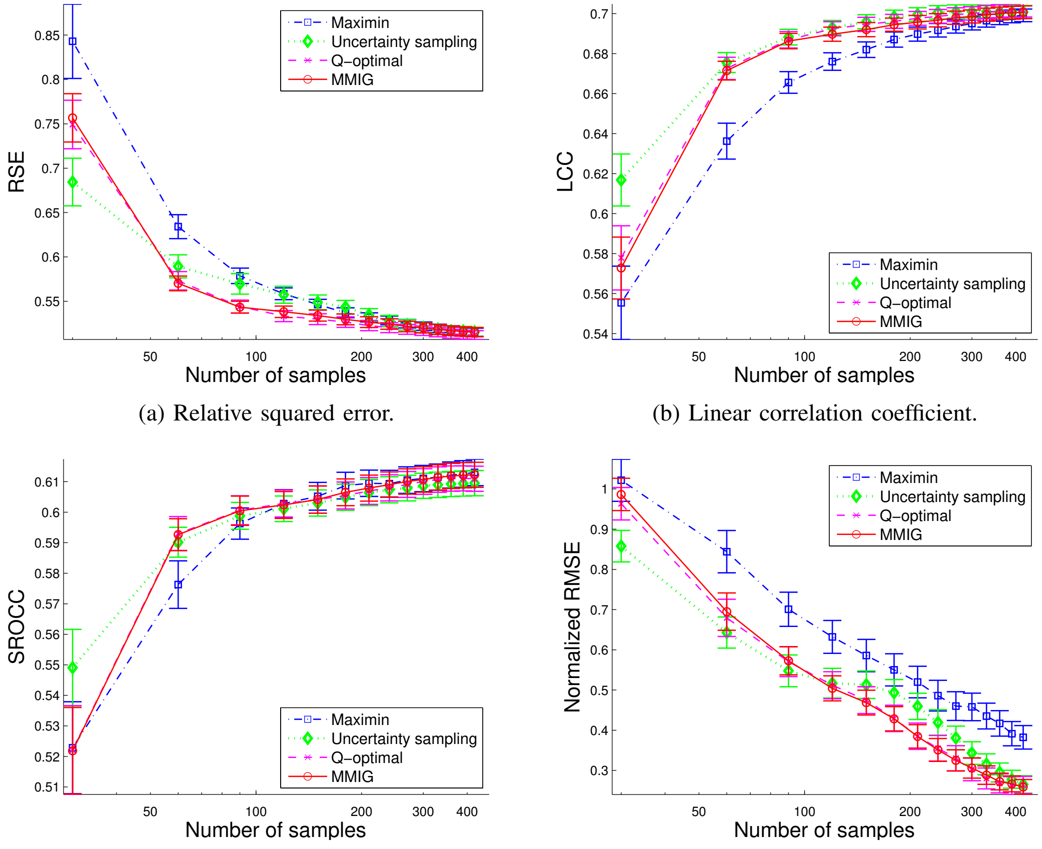
B. Single Stimulus

As in the offline setting, we present one stimulus in each round of rating. To reduce the error caused by the habituation effect in the single-stimulus setting [29], we present a reference video clip to the subject at the beginning of the task. The reference clip has the best quality of all the videos in our experiment. Specifically, its frame rate, resolution, and bit rate are 30 fps, 1920x1080 and 10000 kbps, respectively. Asking a subject to label the same query within a short time can create bias, so each query can be selected only once for each subject.

We evaluate three sampling methods, namely, random, maximin, and Q-optimal sampling. We evaluate Q-optimal first because the method achieves the lowest RSE in our offline experiment. We collect 3,600 samples in 18 trials from MTurk, so each method is adopted in 6 trials.

1) Maximin versus random sampling: We first compare random sampling and space-filling sampling under three regression models: linear regression, random forest regression, and MIQX model. All models are trained on samples collected during a single trial and are tested using samples collected from all other trials. The results shown in Figure 11 confirm that MIQX is preferable to the other regression methods. Moreover, space-filling sampling performs significantly better than random sampling when the MIQX model and linear regression are used. For random forest, space-filling sampling does not outperform random sampling on all metrics, and we believe the reason is that, in contrast to the offline setting, we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2831639, IEEE Transactions on Multimedia



(a) Relative squared error.

(b) Linear correlation coefficient.

(c) Spearman rank-order correlation coefficient.

(d) Normalized root-mean-square error of feature weights.

Fig. 9: Active sampling versus maximin sampling for the MIQX model.

cannot simulate a large number of trials in the online setting to eliminate the impact of outliers.

2) Active versus maximin sampling: Next, we compare active sampling (i.e., Q-optimal sampling) with maximin sampling for our MIQX model by using all the samples collected by random sampling as our testing data. Figure 12 shows that maximin sampling performs significantly better than active sampling based on various metrics, especially RSE. This result is surprising because it contradicts the findings of our offline experiment.

To ensure the performance difference is due to the online setting rather than noise generated by the data collection process, we repeat the offline experiment on newly collected data. As in section V, we divide the data collected by random sampling into two portions: 70% of the data is used for training and 30% of the data is used for testing. Then, different sampling methods are applied to select the next sample from the training data. As shown in Figure 13, the results are similar to those of our offline experiment: active sampling (including Q-optimal and MMIG) performs better than maximin sampling.

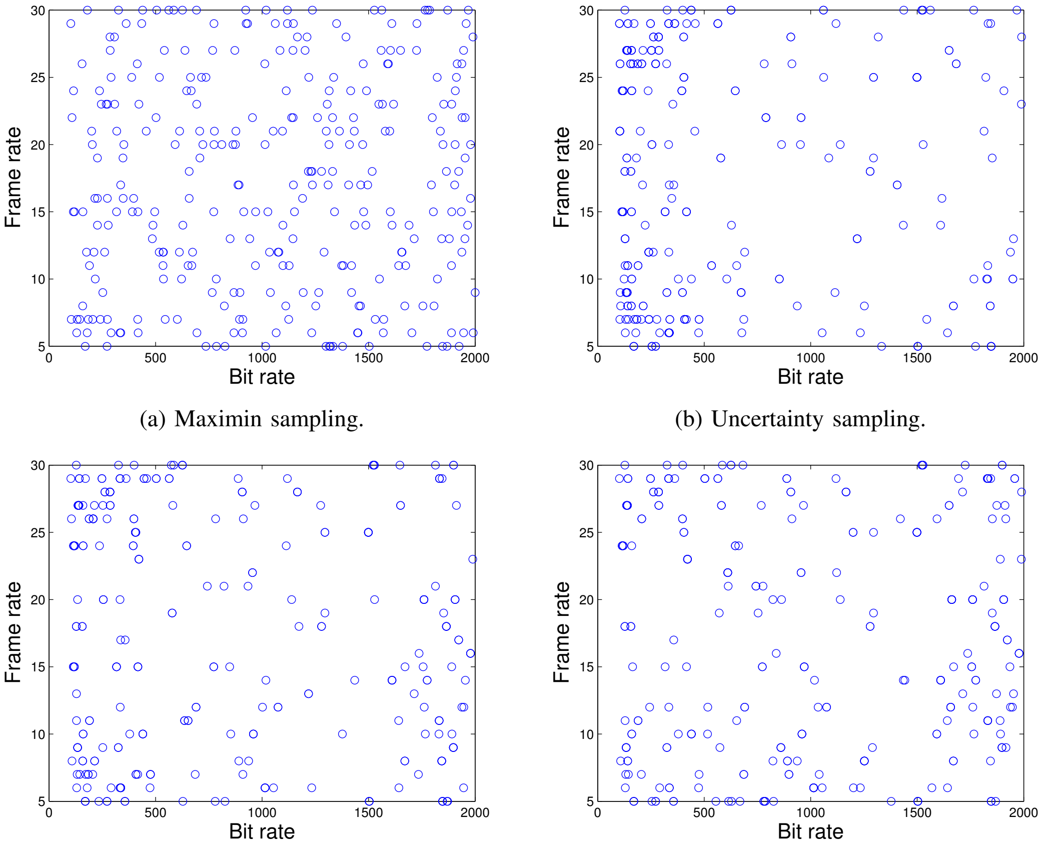
This field experiment reveals the limitations of offline evaluations like those in Section V or previous work [2, 3, 4, 5]. The experiment also demonstrates that, to the best of our knowledge, some inherent difficulties of applying active learning to quality assessment tasks have not been considered. In the following two subsections, we hypothesize and verify two difficulties that degrade the performance of active sampling in our experiment.

3) Difficulty 1 (habituation effect): One important difficulty is the habituation effect (i.e., the contextual effect [29] or memory effect [43]). Subjects tend to give higher scores than usual if they just saw a clip with very bad quality. Because our sampling algorithm is not purely random, the subject might not provide the same scores as they did in random sampling because our algorithm might introduce a systematic bias on the basis of subjects' judgment.

Taking our active sampling as an example, the algorithms tend to select videos with extreme parameter values, as shown in Figure 10. As a result, the subject is very likely to see the alternation between the videos with very good quality and videos with very bad quality, so their scoring standards might be very different from those in an experiment presenting randomly selected videos, even though we attempt to reduce the habituation effect by showing a reference video at the beginning of the experiment.

To verify this hypothesis, we fix our estimation method and use different sets of testing data to perform the evaluation. We train a MIQX model on data obtained by random sampling, and we predict the scores collected from maximin sampling and active sampling based on the model. The results illustrated in Figure 14 show that the MIQX model estimated using data from random sampling can predict scores from maximin

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2831639, IEEE Transactions on Multimedia



(a) Maximin sampling.

(b) Uncertainty sampling.

(c) Q-optimal sampling.

(d) MMIG sampling.

Fig. 10: The first three hundred sample points of different sampling methods for our MIQX model.

sampling much better than it can predict scores from active sampling, especially in terms of the RSE. By contrast, if we assess the prediction performance based on rank coefficients (i.e., SROCC), both evaluations have similar performance. These two results suggest that subjects tend to provide similar scores when they are given videos sampled randomly and uniformly. On the contrary, the order provided by active sampling alters the subjects' standards of rating quality. Although the subjects still rank the video quality similarly (i.e., videos with higher quality receive a better score), the absolute scores are different due to the habituation effect.

4) Difficulty 2 (individual differences): Another important difficulty is inter-subject differences (i.e., each subject has different standards for their judgments). For example, some subjects care more about the smoothness of motion (i.e., prefer high frame rates) while others pay more attention to the quality of each frame (i.e., prefer high bit rates). In our experiment, a subject provides 40 consecutive samples in a trial, so active sampling might try to fit the QoE model of the current subject instead of fitting the average QoE model of the crowd, which is usually the goal of a VQA task.

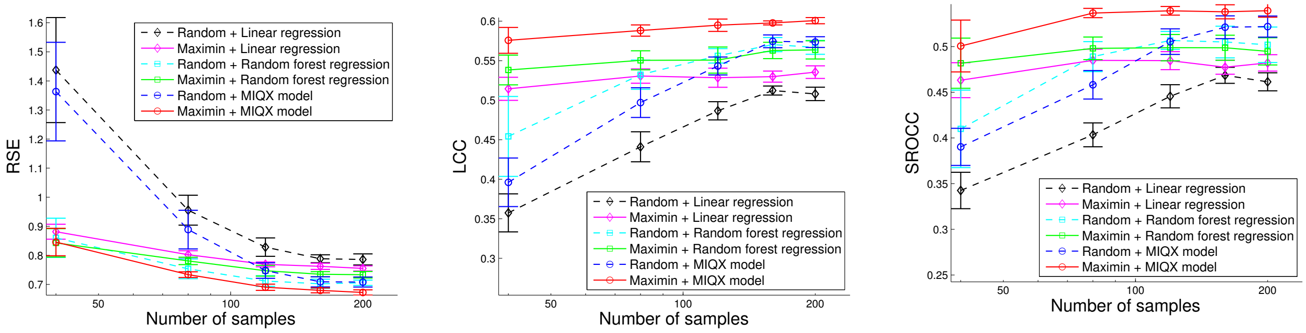
To confirm the effect of inter-subject differences, we compare the performances of two sets of MIQX models. One set is trained and tested using the scores from the same trial, while the other set is trained and tested using the scores from different trials. More specifically, we divide each trial into 10 folds and train a MIQX model using 9 folds of each trial. Next, the performance of the models tested on the remaining 1 fold from the same trial is shown in the rows labeled as ”The same trial” in Table II. Additionally, the other three rows in the table show the performance when testing the models on a fold of all other trials collected by the same sampling method.

Comparison of the performances of random sampling evaluated using different testing data shows that the performance on the same trial is better than the performance on a different trial, especially when the performance is assessed by RSE. Therefore, each subject tends to annotate video quality differently, especially in terms of the absolute scores they provide. Performance differences are also observed in the two other sampling methods. Among all three sampling methods, the performance difference in active sampling is the largest. This result suggests that active sampling fits the QoE model of each individual subject better than it fits the global QoE model.

C. Double Stimulus

To overcome the problems of applying active sampling, we change our testing method from a single stimulus to a double stimulus and propose a hybrid method that combines active sampling and space-filling sampling.

In addition to the compressed video, in the double-stimulus setting, we present a reference clip, which contains minimal compression distortion similar to the reference videos in the single-stimulus experiment. According to [29], the double

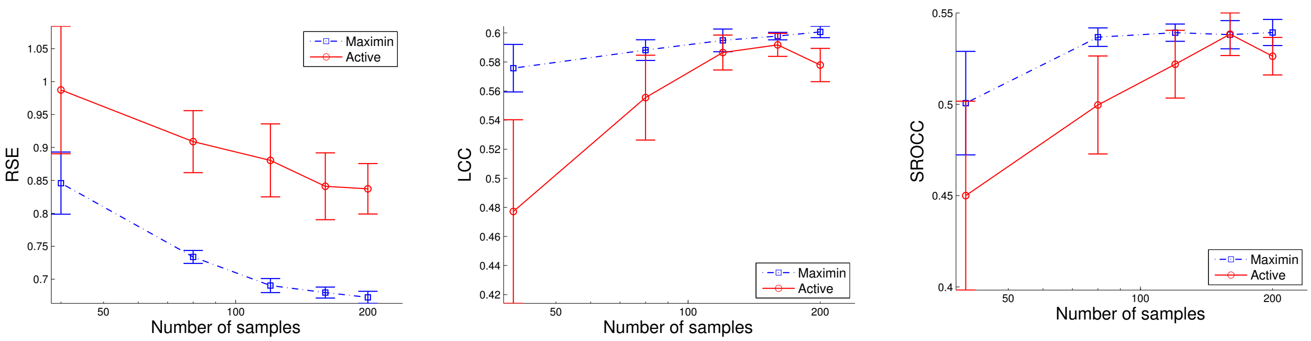


(a) Relative squared error.

(b) Linear correlation coefficient.

(c) Spearman rank-order correlation coefficient.

Fig. 11: Comparison of maximin sampling and random sampling using the single-stimulus setting.

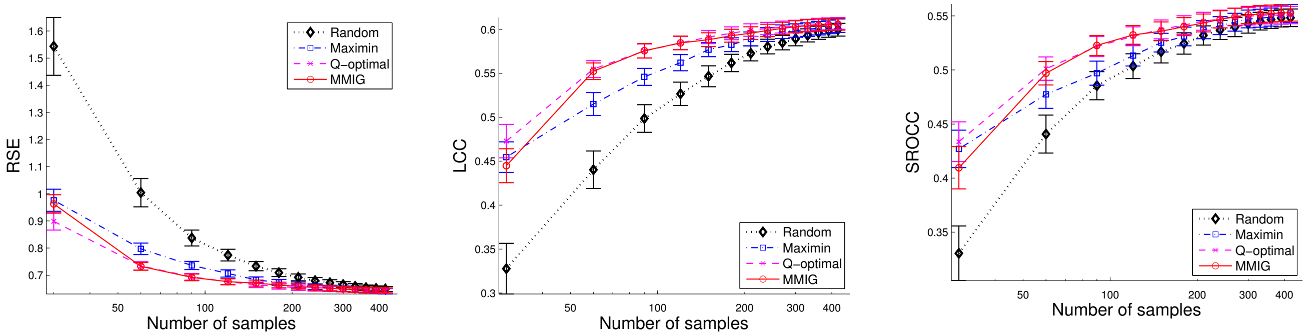


(a) Relative squared error.

(b) Linear correlation coefficient.

(c) Spearman rank-order correlation coefficient.

Fig. 12: Comparison of active sampling and maximin sampling using the single-stimulus setting.



(a) Relative squared error.

(b) Linear correlation coefficient.

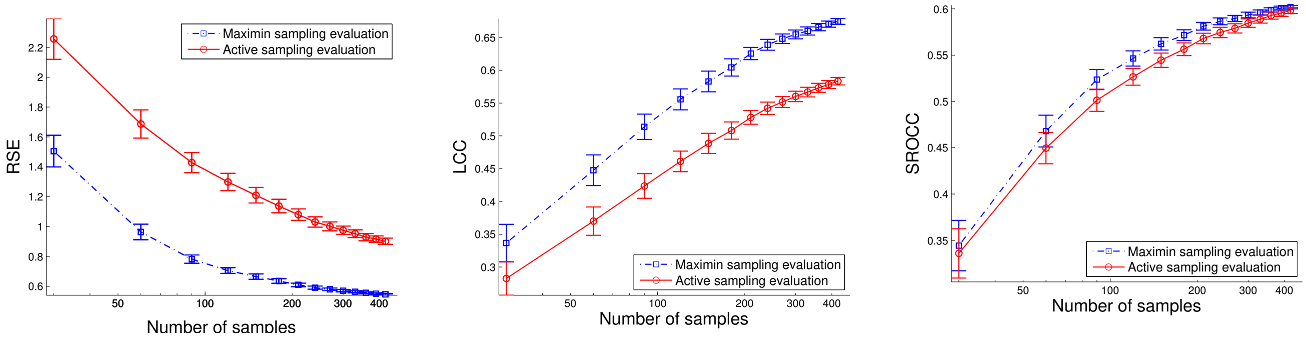
(c) Spearman rank-order correlation coefficient.

Fig. 13: Effects of changing the sampling order of random sampling in our single-stimulus setting.

stimulus is less sensitive to the habituation effect because subjects can judge the quality of the target video by comparison with the quality of the reference video. In our implementation, the full screen of the subject's computer is used to display the first half of one video, and the second half of the other video is shown next. Initially, either video can be shown with a probability of 0.5. After the first half of the first video is complete, we present the second half and then the first half of the other video, and so on.

Figure 14 and Table II, from the experiments in the previous sections, illustrate that space-filling sampling is much less likely to introduce the habituation effect and can effectively reduce the error introduced by inter-subject differences. These experiments motivate us to integrate active sampling and space-filling sampling. We choose MMIG because the samples drawn by MMIG tend to be more evenly distributed than those obtained by Q-optimal, as shown in Figure 10, and this property might be helpful in reducing the habituation effect.

The integration method we adopt is straightforward. We simply add the objective function of MMIG in (13) and the



(a) Relative squared error.

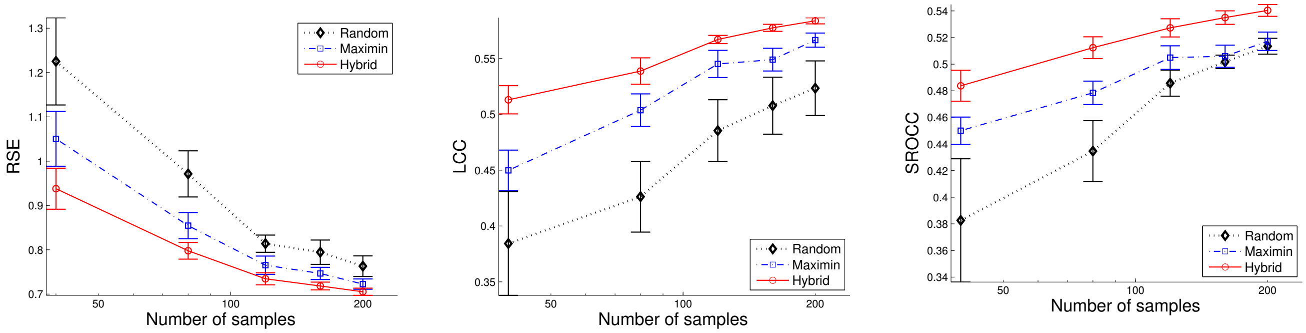
(b) Linear correlation coefficient.

(c) Spearman rank-order correlation coefficient.

Fig. 14: Experimental verification of the habituation effect. The figure uses the annotations collected from active sampling and maximin sampling to evaluate the MIQX model trained by random sampling.

TABLE II: Experimental verification of individual differences. The table shows the accuracy of predicting samples inside and outside the training trial using different sampling methods. The confidence interval is computed with the standard error of the average scores.

Method	Source of test data	RSE	LCC	SROCC
Random Sampling	The same trial	0.734 ± 0.014	0.567 ± 0.008	0.497 ± 0.008
	Different trials	0.839 ± 0.008	0.558 ± 0.003	0.489 ± 0.004
Active Sampling	The same trial	0.643 ± 0.011	0.637 ± 0.007	0.605 ± 0.007
	Different trials	0.742 ± 0.006	0.597 ± 0.003	0.565 ± 0.003
Maximin Sampling	The same trial	0.609 ± 0.012	0.675 ± 0.007	0.570 ± 0.031
	Different trials	0.621 ± 0.006	0.676 ± 0.003	0.563 ± 0.003



(a) Relative squared error.

(b) Linear correlation coefficient.

(c) Spearman rank-order correlation coefficient.

Fig. 15: Comparison of active, maximin and random sampling using the double-stimulus setting.

objective function of maximin in (7) together as

$$\mathbf{x_i^*} = \arg\max_{\mathbf{x_i}\in\mathbf{S}}(U(\mathbf{X}) - U([\mathbf{X};\mathbf{x_i}],[\mathbf{y};y_i]) + \rho\cdot(\min_{\mathbf{x_k}\text{ for }k=1\ldots(i-1)}(d(\mathbf{x_i},\mathbf{x_k})))), \tag{14}$$

where $U(\mathbf{X})$ is defined in (12), d is the distance function used in (7), and ρ is the weight for balancing two goals, which is set to $\frac{\max_{\mathbf{x_i}} U(\mathbf{X})-U([\mathbf{X};\mathbf{x_i}],[\mathbf{y};y_i])}{\max_{\mathbf{x_i}}(\min_{\mathbf{x_k}\text{ for }k=1\ldots(i-1)}(d(\mathbf{x_i},\mathbf{x_k})))}$ in this experiment.

We collect 10 trials for each of the 3 sampling methods, and each trial contains 200 samples labeled by 5 subjects. The testing data for all sampling methods are obtained by random sampling. When evaluating the performance of random sampling, we use all trials other than the training trial as the testing data. The results are presented in Figure 15, which shows that the hybrid method that integrates maximin and MMIG significantly outperforms the alternative solutions in terms of all metrics in the double-stimulus setting.

As analyzed in a previous study [29], double stimulus and single stimulus have their own advantages and disadvantages. For example, single stimulus can be applied to video of arbitrary long length or dynamic quality. On the other hand, double stimulus reduces the noise and bias in the annotations, which is especially appealing in the crowdsourcing setting. Our study shows that the success of active learning approaches for subjective quality assessment tasks strongly depends on the choice of testing method and suggests that the consideration of such deployment issues may be necessary when designing

TABLE III: The first column shows the training sampling strategy and the second row shows the testing sampling strategy. The data are collected in the double-stimulus setting.

Testing / Training	RSE			LCC			SROCC		
	Hybrid	Maximin	Random	Hybrid	Maximin	Random	Hybrid	Maximin	Random
Hybrid	0.501 ± 0.005	0.568 ± 0.005	0.669 ± 0.007	0.710 ± 0.003	0.664 ± 0.004	0.603 ± 0.005	0.677 ± 0.004	0.605 ± 0.004	0.553 ± 0.005
Maximin	0.512 ± 0.004	0.560 ± 0.005	0.644 ± 0.006	0.704 ± 0.003	0.667 ± 0.004	0.606 ± 0.005	0.664 ± 0.004	0.605 ± 0.004	0.550 ± 0.005
Random	0.530 ± 0.005	0.562 ± 0.005	0.642 ± 0.006	0.699 ± 0.003	0.668 ± 0.003	0.603 ± 0.005	0.676 ± 0.004	0.611 ± 0.004	0.552 ± 0.005

new active learning algorithms.

VII. Limitations

Figure 15 evaluates the sampling strategies for only the first 200 samples. To assess the long-term performance of each sampling method, we train our MIQX model on all 2,000 samples collected by one sampling method and evaluate the model using samples collected from other sampling methods in the double-stimulus setting. In addition, we perform 10-fold cross-validation on all 2,000 samples collected by each sampling method (i.e., training and testing using the same sampling strategy).

The results are shown in Table III. The RSE of the model trained by the hybrid sampling and tested by random sampling shows that the performance is slightly worse than that of the model trained by maximin sampling and tested by random sampling. Therefore, even though we use double stimulus and a hybrid method, our active learning algorithm still produces some small bias.

Another limitation is about our bitrate setup. In the experiments, we set our bitrate from 100 to 2,000 kbps in Section V and from 100 to 2,500 kbps in Section VI. The quality in this bitrate range is relatively low compared with the services from popular online video streaming platforms nowadays (e.g., YouTube encodes live stream 1080p videos using 3,000–6,000 kbps in 2018).

One of the main reasons behind our bitrate range choice is to support prospective studies about tuning video hyper-parameter dynamically in a low bandwidth connection channel. Furthermore, the online workers might not have the adequate skills or hardware to identify the subtle difference among high-quality videos, which might introduce severe noise into the dataset. As we can see in Figure 4b, it is already hard for annotators to tell the difference between videos with 1,000 kbps and the ones with 2,000 kbps.

As the infrastructure of Internet keeps improving, the quality expectation from the general public will also increase. In response to the trend, the bitrate of focus might need to cover a larger interval (e.g., 100 to 6,000 kbps). The change will make modeling QoE under low bitrate more difficult, especially for random sampling strategy, given the same budget, but it also makes such modeling less important. How to develop and evaluate active learning and crowdsourcing methods to cope with the continuous shift of the mainstream bitrate range requires further investigation.

VIII. Lesson Learned and Future Work

Modeling the QoE of multimedia applications often requires the consideration of a large variety of QoS factors. Appropriate sampling methods are required to cope with the large parameter space. In general, several key challenges and issues arise for multidimensional QoE testing and modeling.

- Efficiency: Crowdsourcing provides a powerful tool to conduct subjective experiments with a large pool of subjects; nevertheless, appropriate sampling strategies, which lead to accurate QoE models without incorporating undesirable side effects as a result of the sampling strategy, are required for efficient data collection.
- Habituation effect: The QoE score is affected by the order of sampling because the quality expectation of subjects decreases after seeing multimedia with poor quality.
- Subject bias and individual differences: User (rating) diversity leads to different standards for the subjects' judgments. Although most of the noise can be eliminated after collecting responses from a sufficient number of subjects, it is difficult to acquire unbiased scores for active sampling methods in the early stages of the annotation collection process.

In our experiments, we vary the sampling strategies (random, maximin, active sampling: uncertainty, Q-optimal, and MMIG), the regression models (linear, kernel, random forest regression, and MIQX model), and the testing methodology (single stimulus and double stimulus) to analyze their impact on the efficiency and accuracy of multidimensional QoE modeling. The following lessons can be derived from the experiments as key take-aways for the readers.

- Maximin sampling obtains even coverage of the parameter space and usually significantly outperforms random sampling, especially when only few samples are collected.
- Uncertainty sampling tends to sample only points near the boundary of the feature space, which degrades its performance, while MMIG and Q-optimal could alleviate the problem.
- Field experiments demonstrate that active learning algorithms, which perform very well in offline experiments (simulating the sampling order), can perform worse than passive sampling (e.g., maximin sampling) due to the habituation effect and individual differences among annotators.
- An appropriate testing methodology must be carefully selected when active learning approaches are adopted.
  - Active learning with a double-stimulus test design may overcome the habituation effect.
- Active learning in multidimensional QoE modeling requires further investigation.
  - To study the impact of (triggered and even intensified) habituation effects in greater detail;
  - To overcome individual differences.

Some general recommendations for crowdsourcing multidimensional QoE modeling are given below.

- Test active learning in field experiments not only in an offline setting. Changing the sampling order may introduce hidden influence factors for subjects that are not captured in the offline setting.

- When designing an active learning algorithm for subjective quality assessment, consider the solutions for the habituation effect and individual differences at the beginning. In the experiments, we show that reference videos and mixing active sampling with space-filling sampling are effective techniques. Other potential solutions include:
  - Taking into account the previous QoE scores as additional features in the regression problem [43];
  - Modeling the user diversity or identifying different user groups [44, 45];
  - Providing additional training for subjects [22, 23];
  - Integrating proper reliability questions into the task design to filter unreliable users (e.g., Hofeld et al. [13]).
- Remove subject bias and align the scores from different subjects. All the scores labeled by an individual subject are shifted such that the average score (for each subject) is 4, as suggested by Janowski and Pinson [39].
- In general, it is difficult for online workers to distinguish the subtle quality differences among high-quality videos. More sophisticated noise removal techniques need to be considered when videos with high bitrate are compared using crowdsourcing.
- Non-linear interaction between QoS factors strongly influence QoE. For instance, increasing frame rate increases the quality if bitrate is high, but decrease the quality if bitrate is low. When modeling high-order interactions between multiple QoS factors and QoE, consider the MIQX model, which is simple, probabilistically interpretable, and performs as good as more complicated models, such as random forest.

IX. Conclusion

Modeling the relationships between multiple QoS parameters and QoE used to be an expensive task because an accurate model requires a large number of annotated samples. To improve the process, we extend the IQX model and devise active sampling algorithms for our multidimensional IQX (MIQX) model.

Our experiments verify that the proposed MIQX model is effective for VQA tasks, and we find that maximin sampling almost always performs better than random sampling. We also demonstrate that offline experiments are not sufficient to verify the practical effectiveness of active sampling approaches because the habituation effect and individual differences strongly influence the performance of QoE modeling.

Finally, our field experiments show that combining maximin and MMIG sampling reduces the number of samples required in the double-stimulus setting, and we suggest that solutions for the habituation effect and individual differences need to be investigated when applying active learning to subjective quality assessment in future work.

X. Acknowledgement

This material is based upon work supported in part by DFG Crowdsourcing (HO 4770/2-2). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

[1] M. Fiedler, T. Hoßfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service." *IEEE Network*, pp. 36–41, 2010. 1, 2, 3, 6

[2] V. Menkovski, G. Exarchakos, and A. Liotta, "Tackling the sheer scale of subjective QoE," in *International Conference on Mobile Multimedia Communications*. Springer, 2011, pp. 1–15. 1, 3, 6, 10

[3] V. Menkovski and A. Liotta, "Adaptive psychometric scaling for video quality assessment," *Signal Processing: Image Communication*, vol. 27, no. 8, pp. 788–799, 2012. 1, 3, 6, 10

[4] P. Ye and D. Doermann, "Active sampling for subjective image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4249–4256. 1, 3, 6, 10

[5] B. Osting, J. Xiong, Q. Xu, and Y. Yao, "Analysis of crowdsourced sampling strategies for hodgerank with sparse random graphs," *Applied and Computational Harmonic Analysis*, vol. 41, no. 2, pp. 540–560, 2016. 1, 3, 6, 10

[6] T. Wang, A. Pervez, and H. Zou, "VQM-based QoS/QoE mapping for streaming video," in *Broadband Network and Multimedia Technology (IC-BNMT), 2010 3rd IEEE International Conference on*. IEEE, 2010, pp. 807–812. 1, 2

[7] M. Michalos, S. Kessanidis, and S. Nalmpantis, "Dynamic adaptive streaming over HTTP," *Journal of Engineering Science and Technology Review*, vol. 5, no. 2, pp. 30–34, 2012. 1

[8] ITU-T COM 9-C-60-E, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II (FR-TV2)," 2003. 1, 4, 5

[9] M. Elkotob, D. Grandlund, K. Andersson, and C. Ahlund, "Multimedia QoE optimized management using prediction and statistical learning," in *Local Computer Networks (LCN), 2010 IEEE 35th Conference on*. IEEE, 2010, pp. 324–327. 1, 2, 3

[10] M. S. Mushtaq, B. Augustin, and A. Mellouk, "Empirical study based on machine learning approach to assess the QoS/QoE correlation," in *Networks and Optical Communications (NOC), 2012 17th European Conference on*. IEEE, 2012, pp. 1–7. 1, 2, 6, 7

[11] F. Battisti, M. Carli, and P. Paudyal, "QoS to QoE mapping model for wired/wireless video communication," in *Euro Med Telco Conference (EMTC), 2014*. IEEE, 2014, pp. 1–6. 1, 3

[12] T. Hoßfeld, L. Skorin-Kapov, P. E. Heegaard, M. Varela, and K.-T. Chen, "On additive and multiplicative QoS-QoE models for multiple QoS parameters," in *Proceedings of the 5th ISCA/DEGA Workshop on Perceptual Quality of Systems PQS 2016*. ISCA, 2016. 1, 2

[13] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014. 1, 3, 15

[14] K.-T. Chen, C.-J. Chang, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "Quadrant of euphoria: a crowdsourcing platform for QoE assessment," *IEEE Network*, vol. 24, no. 2, 2010. 1, 3

[15] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *Multimedia (ISM), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 494–499. 1, 2, 3

[16] S. Aroussi and A. Mellouk, "Survey on machine learning-based QoE-QoS correlation models," in *Computing, Management and Telecommunications (ComManTel), 2014 International Conference on*. IEEE, 2014, pp. 200–204. 2

[17] M. Alreshoodi and J. Woods, "Survey on QoE\QoS correlation models for multimedia services," *International Journal of Distributed and Parallel Systems*, pp. 53–72, 2013. 2

[18] R. Schatz, T. Hoßfeld, L. Janowski, and S. Egger, "From packets to people: Quality of experience as a new measurement challenge," in *Data traffic monitoring and analysis*. Springer, 2013, pp. 219–263. 2

[19] D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "A survey on parametric QoE estimation for popular services," *Journal of Network and Computer Applications*, vol. 77, pp. 1–17, 2017. 2

[20] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE transactions on broadcasting*, vol. 57, no. 2, pp. 165–182, 2011. 2

[21] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2015. 2

[22] T. Hoßfeld and C. Keimel, "Crowdsourcing in QoE evaluation," in *Quality of Experience*. Springer, 2014, pp. 315–327. 3, 15

[23] B. Gardlo, S. Egger, and T. Hoßfeld, "Do scale-design and training matter for video QoE assessments through crowdsourcing?" in *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia*. ACM, 2015, pp. 15–20. 3, 15

[24] M. Seufert, O. Zach, T. Hoßfeld, M. Slanina, and P. Tran-Gia, "Impact of test condition selection in adaptive crowdsourcing studies on subjective

quality," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, pp. 1–6. 3, 5

[25] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, no. 55-66, p. 11, 2010. 3, 4, 5

[26] J. Korhonen, N. Burini, J. You, and E. Nadernejad, "How to evaluate objective video quality metrics reliably," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. IEEE, 2012, pp. 57–62. 3

[27] T. Hoßfeld, C. Moldovan, and C. Schwartz, "To each according to his needs: Dimensioning video buffer for specific user profiles and behavior," in *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*. IEEE, 2015, pp. 1249–1254. 3

[28] T. F. Coleman and Y. Li, "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM Journal on optimization*, vol. 6, no. 2, pp. 418–445, 1996. 4

[29] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies." in *VCIP*, 2003, pp. 573–582. 4, 9, 10, 11, 13

[30] L. Pronzato and W. G. Müller, "Design of computer experiments: space filling and beyond," *Statistics and Computing*, vol. 22, no. 3, pp. 681–701, 2012. 4, 5

[31] R. M. Castro, R. Willett, and R. D. Nowak, "Faster rates in regression via active learning." in *NIPS*, 2005, pp. 179–186. 4

[32] M. E. Johnson, L. M. Moore, and D. Ylvisaker, "Minimax and maximin distance designs," *Journal of statistical planning and inference*, vol. 26, no. 2, pp. 131–148, 1990. 4

[33] K. Yu, J. Bi, and V. Tresp, "Active learning via transductive experimental design," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 1081–1088. 5

[34] D. J. C. MacKay, "Information-based objective functions for active data selection." *Neural Computation*, pp. 590–604, 1992. 5

[35] V. V. Fedorov, *Theory of optimal experiments*. Elsevier, 1972. 5

[36] "Big buck bunny movie." (last access: July 2015). [Online]. Available: http://www.bigbuckbunny.org 6

[37] "Tears of steel movie." (last access: July 2015). [Online]. Available: https://mango.blender.org/ 6

[38] "X264." (last access: July 2015). [Online]. Available: http://www.videolan.org/developers/x264.html 6

[39] L. Janowski and M. H. Pinson, "Subject bias: Introducing a theoretical user model." in *QoMEX*, 2014, pp. 251–256. 7, 15

[40] B. Lakshminarayanan and Y. W. Teh, "Inferring ground truth from multi-annotator ordinal data: a probabilistic approach," *arXiv preprint arXiv:1305.0015*, 2013. 7

[41] A. Azzalini and A. W. Bowman, "Applied smoothing techniques for data analysis," *Oxford Statistical Science Series, Oxford*, 1997. 7

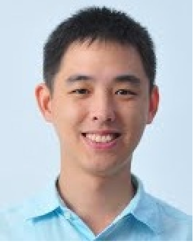
[42] L. Breiman, "Random forests," *Machine learning*, 2001. 7

[43] T. Hoßfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler, "The memory effect and its implications on web qoe modeling," in *Teletraffic Congress (ITC), 2011 23rd International*. IEEE, 2011, pp. 103–110. 10, 15

[44] T. Hoβfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*. IEEE, 2011, pp. 131–136. 15

[45] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, "QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS," *Quality and User Experience*, vol. 1, no. 1, p. 2, 2016. 15

Haw-Shiuan Chang received his B.S. degree in Electrical Engineering and Computer Science from National Chiao Tung University, Hsinchu, Taiwan in 2011. He was a research assistance in Research Center for Information Technology Innovation and Institute of Information Science, Academia Sinica, Taipei, Taiwan from 2012 to 2015. He is currently a MS/PhD student at UMass Amherst, and his research interests include natural language processing, clustering, and active learning.



Chih-Fan Hsu is a research assistant in Academia Sinica, Taiwan, from 2014, and a Ph.D. candidate in department of electrical engineering in National Taiwan University, Taipei, Taiwan, from 2018. His research interests cover computer vision, machine learning, and quality of experience.

Tobias Hoßfeld is professor at the Chair of Communication Networks at the University of Würzburg, Germany, since 2018. He finished his PhD in 2009 and his professorial thesis in 2013 at the University of Würzburg. From 2014 to 2018, he was head of the Chair "Modeling of Adaptive Systems" at the University of Duisburg-Essen, Germany. He published more than 100 research papers and received the Fred W. Ellersick Prize 2013 (IEEE Communications Society).

Kuan-Ta Chen (a.k.a. Sheng-Wei Chen) is s a Research Fellow at the Institute of Information Science and the Research Center for Information Technology Innovation (joint appointment) of Academia Sinica. He is currently the Chairman of Taiwan Data Science Association, the Director of Artificial Intelligence Foundation, the Director of Taiwan AI Academy, and the CTO of E.SUN Financial Holding Company. He was an Associate Editor of IEEE Transactions on Multimedia (IEEE TMM) during 2011 to 2014 and has been an Associate Editor of ACM Transactions on Multimedia Computing, Communications, and Applications (ACM TOMM) since 2015. He organized ACM Multimedia Systems 2017 in Taiwan and served the lead program chair of ACM Multimedia 2017. He is a Senior Member of ACM and a Senior Member of IEEE..