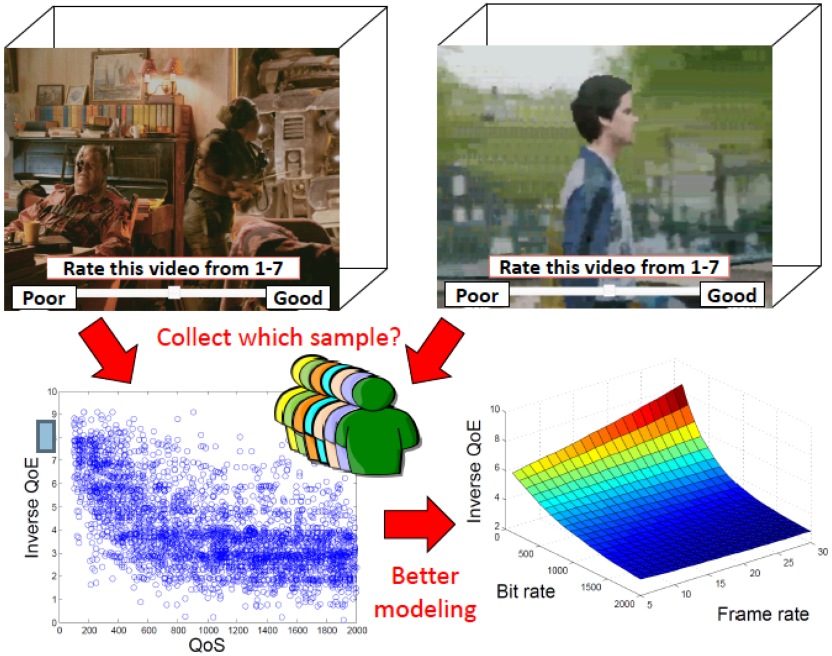
Introduction

- As increasing multimedia content becomes available, the optimization of users' experiences on multimedia given limited resources is more important
- Challenges
  - Unknown mechanism for human to judge the quality -> expensive process of collecting subjects' opinions is usually required for satisfactory QoE estimation
  - There are many dynamic QoS factors that affect QoE in users' minds, e.g., bitrate, resolution, delay, …

QoE Modeling

- Standard approach
  1) random (grid) sampling in QoS space
  2) subjects are asked to score on those samples
  3) modeling the relationship between QoE and QoS

- Goal of this article
  - actively select (informative) samples to better model the relationships between QoS parameters and QoE with fewer samples



Multidimensional IQX (MIQX) Modeling

- Goal: predict the QoE based on QoS

$$y = f(\mathbf{x})$$

- IQX model

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

- Multidimensional IQX model (MIQX)

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

$$\theta = [\alpha \;\; \gamma \;\; \mathbf{w}]$$

Training Multidimensional IQX (MIQX) Model

- min 2-norm errors between f(x_i) and y_i for all i

$$E(\theta, \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \left(f_\theta(\mathbf{x_i}) - y_i\right)^2 = \left(\mathbf{y} - F_\theta(\mathbf{X})\right)^T \left(\mathbf{y} - F_\theta(\mathbf{X})\right)$$

$$= \left(\alpha \cdot e^{-\Phi(\mathbf{X},\mathbf{w})} + \gamma \cdot \mathbb{1} - \mathbf{y}\right)^T \left(\alpha \cdot e^{-\Phi(\mathbf{X},\mathbf{w})} + \gamma \cdot \mathbb{1} - \mathbf{y}\right),$$

  - Valid range $\Theta = \{[\alpha, \gamma, \mathbf{w}] \mid \mathrm{QoE}_{\min} \leq \gamma \leq \mathrm{QoE}_{\max} \wedge 0 \leq \alpha \leq (\mathrm{QoE}_{\max} - \mathrm{QoE}_{\min})\},$

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

Adaptive Sampling for QoE Modeling

- Design the sample presentation order such that they can reduce the number of samples required to build an accurate model
  - Grid and Random Sampling
  - Online Space-filling Sampling
  - Active Sampling

Grid and Random Sampling

- Uniform grid sampling
  - easy to implement
  - requires users to set the number of samples in advance (the number cannot be arbitrary)
- Random sampling
  - randomly and uniformly acquires the next sample
  - some large areas in the sampling space may not covered by any sample when the budget is insufficient
  - wastes the annotation in some cases, e.g., two very similar consecutive samples
  - biased sampling results for a subject (e.g., many more high-quality videos compared to low-quality videos)

Online Space-Filling Sampling

- Maximin sampling
  - Select the i-th sample x_i as farther from the chosen samples **x** as possible

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

  - Maximin sampling tends to acquire samples near the boundaries of valid range initially

Active Sampling

- Select the next sample that is most informative for estimating the model parameters
- Information estimation: probabilistic MIQX model
  - Error: normal $\mathcal{N}(0, \sigma_\nu)$
  - α, γ are uniform within their range
  - *w* is Gaussian with a mean and covariance matric of 0 and $\frac{1}{\lambda}I$

$$P(\theta|\mathbf{y}, \mathbf{X}) \propto P(\mathbf{y}|\theta, \mathbf{X})P(\theta|\mathbf{X})$$
$$\propto e^{-(\mathbf{y}-F_\theta(\mathbf{X}))^T(\mathbf{y}-F_\theta(\mathbf{X}))} e^{-\lambda \mathbf{w}^T \mathbf{w}}$$

Active Sampling

- Uncertainty sampling
  - sample the most uncertain point for the current model in the feature space
  - uncertainty: the variance of the prediction of the current QoE-QoS model $\sigma^2(f_\theta(\mathbf{x}))$

$\Rightarrow$ select $\arg\max_{\mathbf{x}\in S} \sigma^2(f_\theta(\mathbf{x}))$ to minimize the overall prediction variance

Active Sampling

- Sample with the highest prediction variance is usually the sample on the edge of the valid feature space -> suffer from outlier more easily
- focus on minimizing the uncertainty of the prediction from highly probable **x** (considering P(x))
  - Minimizing prediction variance -> Q-optimal
  - Maximizing the information gain
    -> mean marginal information gain (MMIG)

Prob. of observing
QoS parameters

Q-Optimal

- Minimize the variance weighted by the feature distribution

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

  - Next sample

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

MMIG

- Minimize the uncertainty of the prediction probability distribution $P(f_\theta(\mathbf{x_u}))$
  - average uncertainty over all valid features **x_u** on the basis of entropy

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

  - next sample

$$\mathbf{x_i^*} = \arg\max_{\mathbf{x_i}\in\mathbf{S}} \left(U(\mathbf{X},\mathbf{y}) - U([\mathbf{X};\mathbf{x_i}],[\mathbf{y};y_i])\right) = \arg\max_{\mathbf{x_i}\in\mathbf{S}}$$

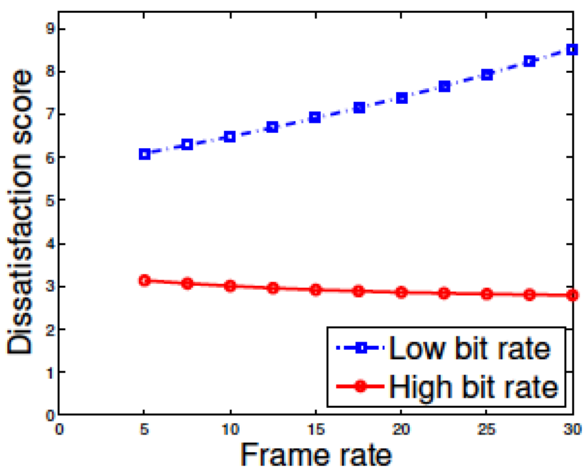
$$(-\frac{1}{2}\int_{\mathbf{x_u}} P(\mathbf{x_u}) \log\left(1 - \frac{g(\mathbf{x_i})^T A^{-1} g(\mathbf{x_u})}{\sigma^2(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x_u}))(\sigma^2(f_{\theta(\mathbf{y},\mathbf{X})}(\mathbf{x_i})) + \sigma_\nu^2)}\right) d\mathbf{x_u})$$

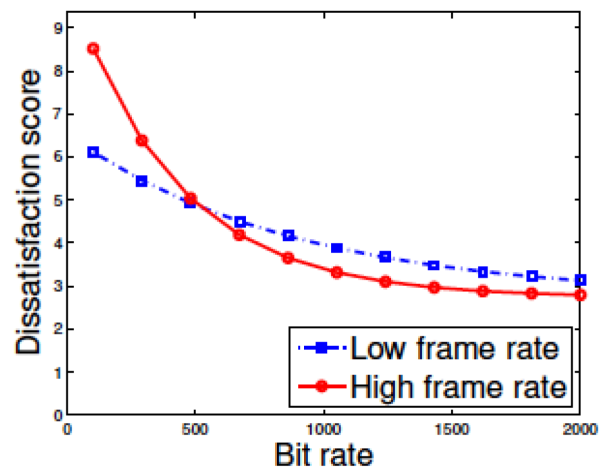
Performance Evaluation

1. Collect QoE scores from video clips with randomly selected QoS parameters
2. Change the collection order offline to evaluate the sampling methods

Experiment Design

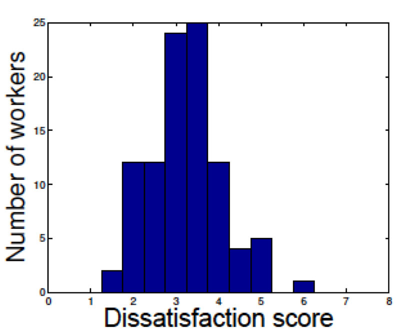
- Video characteristics: bitrate, frame rate, resolution, temporal complexity, and spatial complexity
- Normalized feature space [0,1]
- Use inverse of QoE scores as the prediction target of the regression task: dissatisfaction score
- Interaction between features exist: add 2nd-order interaction terms to the model



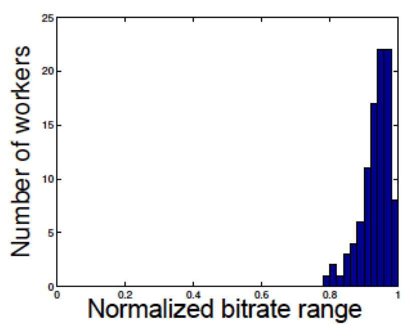


Dataset

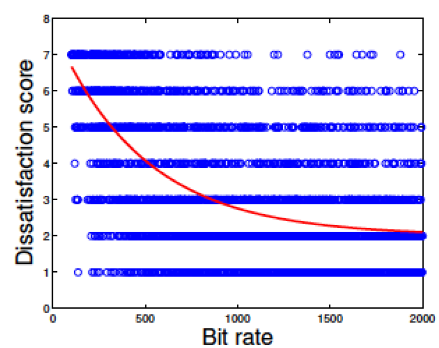
- 3318 annotations from 97 subjects using Amazon Mechanical Turk (MTurk) and Bounty Worker
  - 7-level scale
- 10-second H264 video randomly chosen from Big Buck Bunny and Tears of Steel
  - Bitrate: [100, 2000] kbps
  - Frame rate: [5,30] fps
  - Resolution: {480, 600, 720, 840, 960, 1080} height
- Severe subject bias → normalization



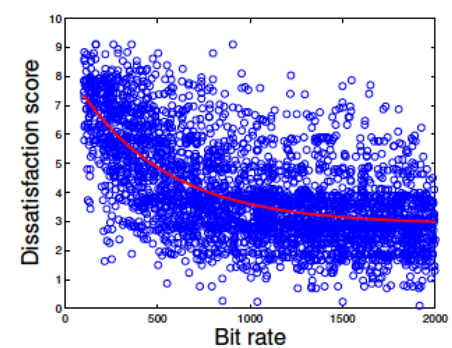
(a) Distribution of average scores.



(b) Distribution of the normalized bit rate range.



(a) Before bias removal.



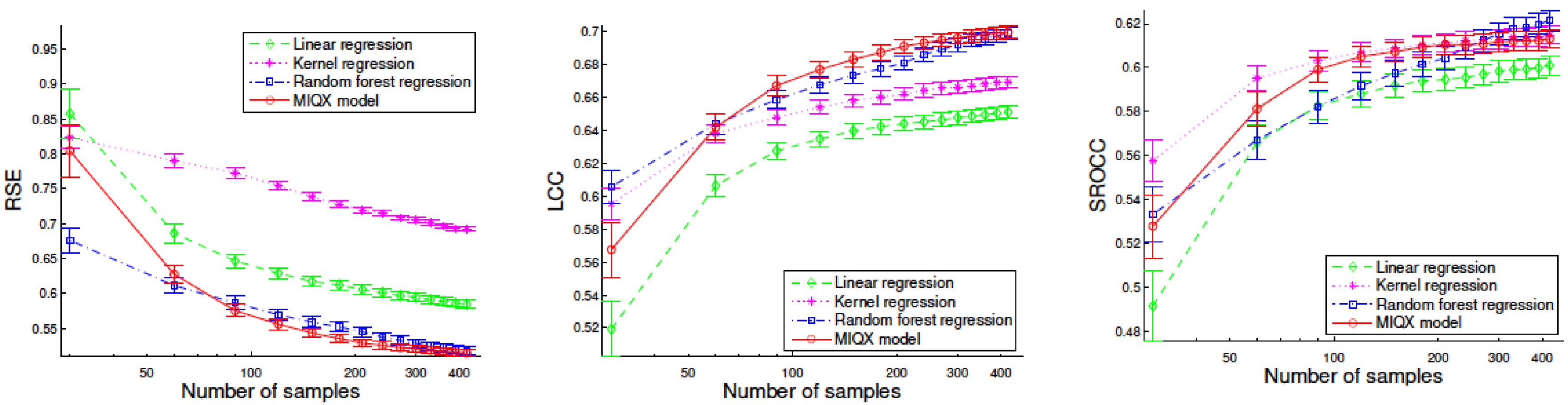
(b) After bias removal.

Evaluation Sampling Methods

- Conduct 200 trials and make each method collect different samples in each trial by injecting some randomness into the sampling process
  - 70% for training and 30% for testing
  1. Randomly select 10 sample from the training pool
  2. Let the method choose the next query
- Evaluation
  - Prediction accuracy: similarity between the prediction and the annotations in testing pool
    - relative squared error (RSE), linear correlation coefficient (LCC), and Spearman rank-order correlation coefficient (SROCC)
  - Parameter accuracy: RMSE of w

Regression Models

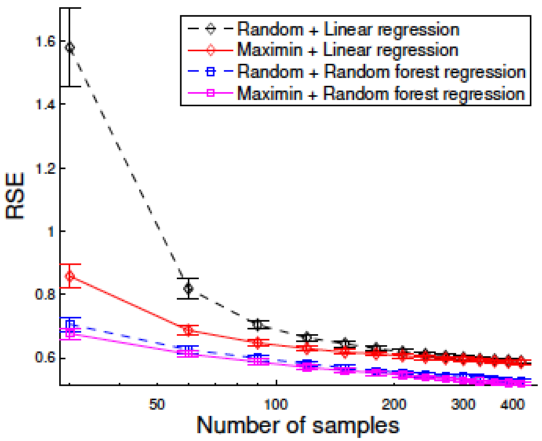
- MIQX (with 2nd-order interaction terms)
- Linear regression (with 2nd-order interaction terms)
- Nadaraya-Watson kernel regression with Gaussian kernel
- Random forest



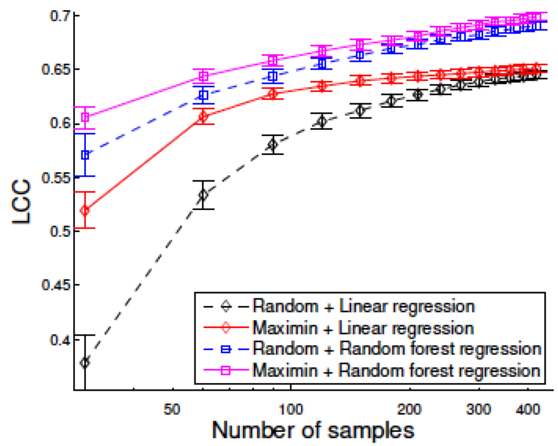
(a) Relative squared error. (b) Linear correlation coefficient. (c) Spearman rank-order correlation coefficient.

Maximin vs. Random Sampling

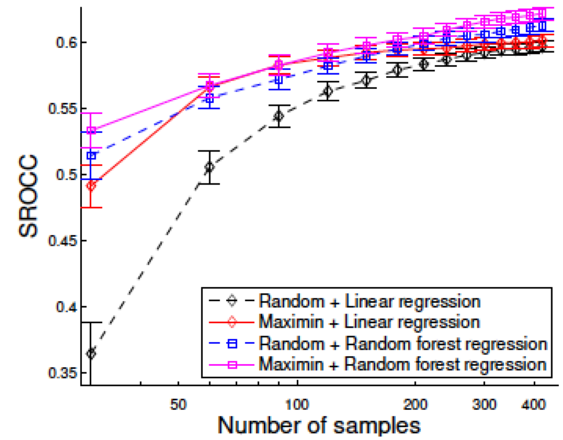
- Maximin sampling leads to more accurate model



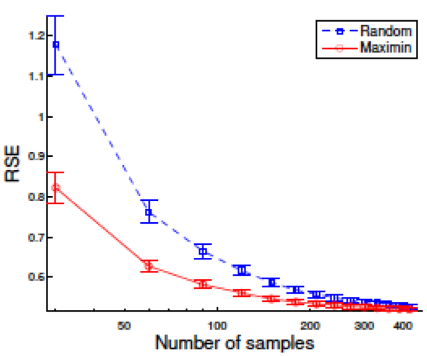
(a) Relative squared error.



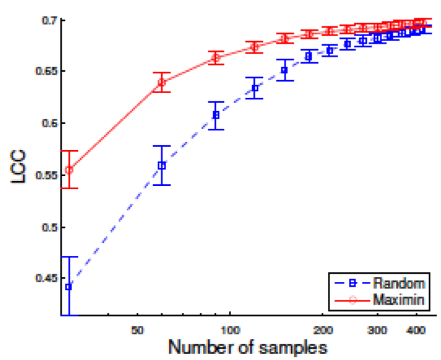
(b) Linear correlation coefficient.



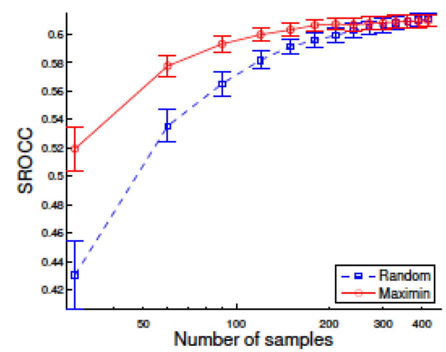
(c) Spearman rank-order correlation coefficient.



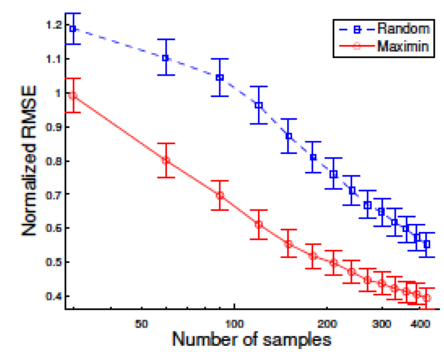
(a) Relative squared error.



(b) Linear correlation coefficient.



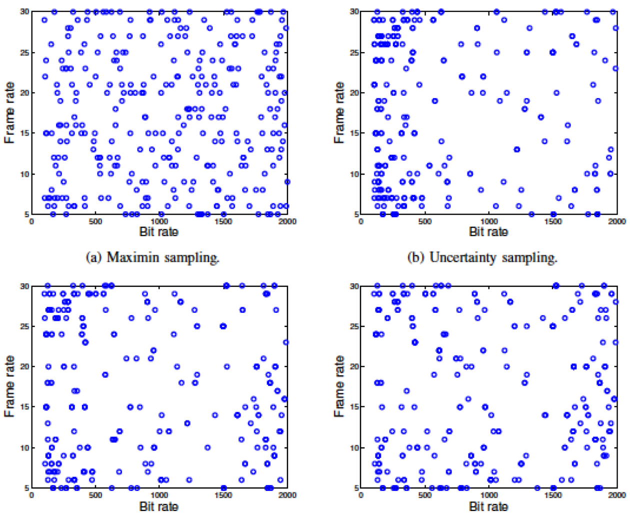
(c) Spearman rank-order correlation coefficient.



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

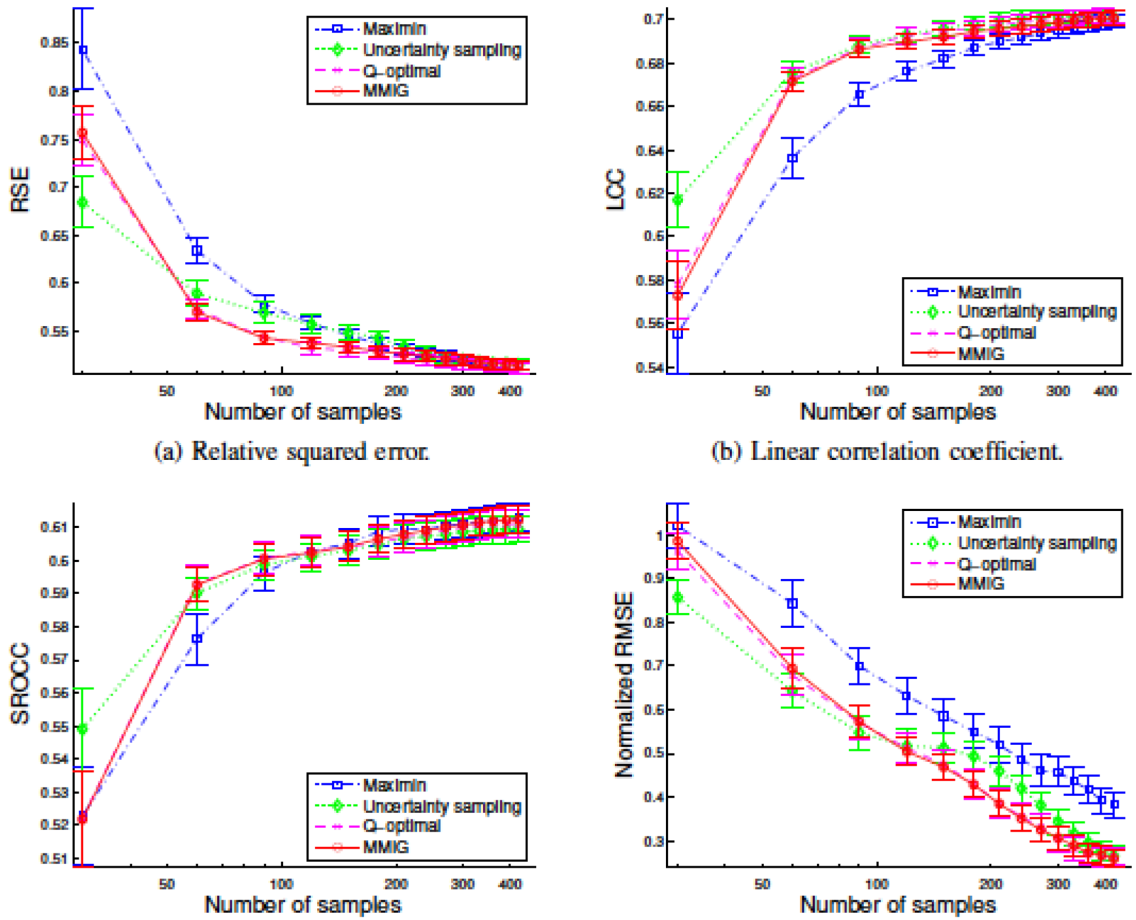
Active vs. Maximin Sampling

- For MIQX model



(a) Maximin sampling. (b) Uncertainty sampling.

(c) Q-optimal sampling. (d) MMIG sampling.



(a) Relative squared error. (b) Linear correlation coefficient.

(c) Spearman rank-order correlation coefficient. (d) Normalized root-mean-square error of feature weights.

Field Experiment for Realistic Online Setting

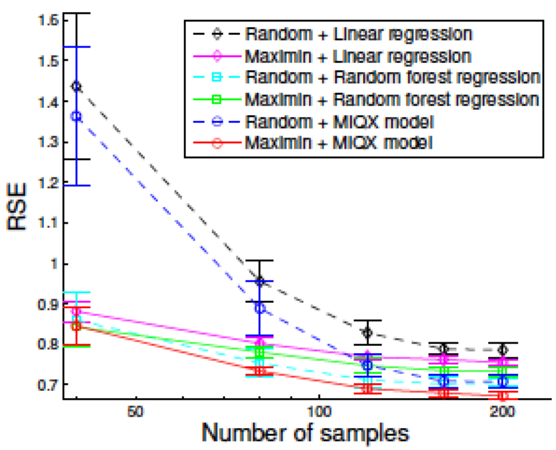
- Experiment design
  - Several trials for each sampling method
  - Randomly assign each subject to a trial
  - In each trail, the query for each subject is determined online based on the previous queries
    - Each subject rate 40 samples
  - 5 subjects (200 samples) collected for each trail
  - The first 10 queries for each subject are randomly selected
  - Shift scores based on the updated average score for bias removal
  - Uniformly sample 3000 QoS parameters (2500 kbps)

Single Stimulus

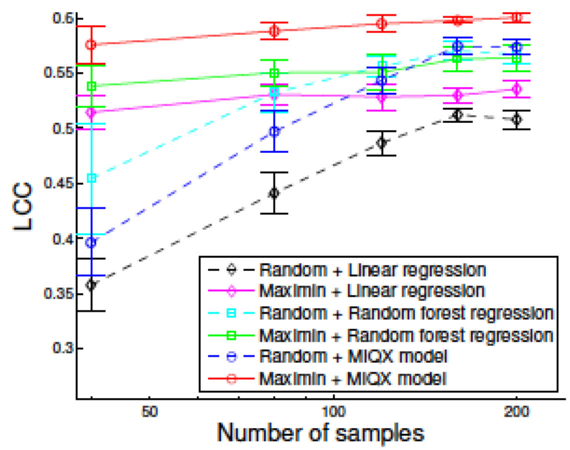
- One stimulus in each round of rating
- The reference video clip is shown to the subject at the beginning of the task (10000 kbps, 1080p, 30 fps)
- Methods: random, maximin, and Q-optimal
- 3600 samples, 18 trials (6 trials for each method)

Maximin vs. Random Sampling

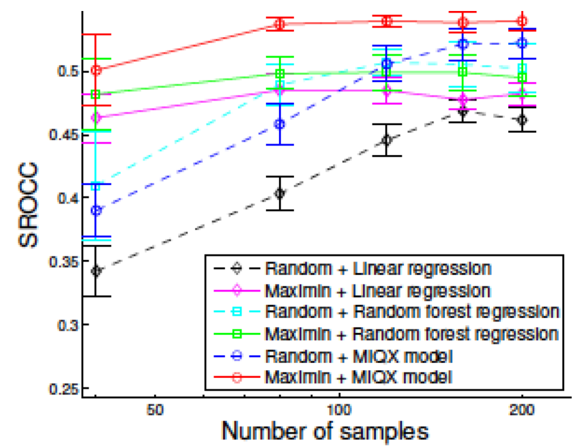
- MIQX > others, maximin > random (except RF)



(a) Relative squared error.



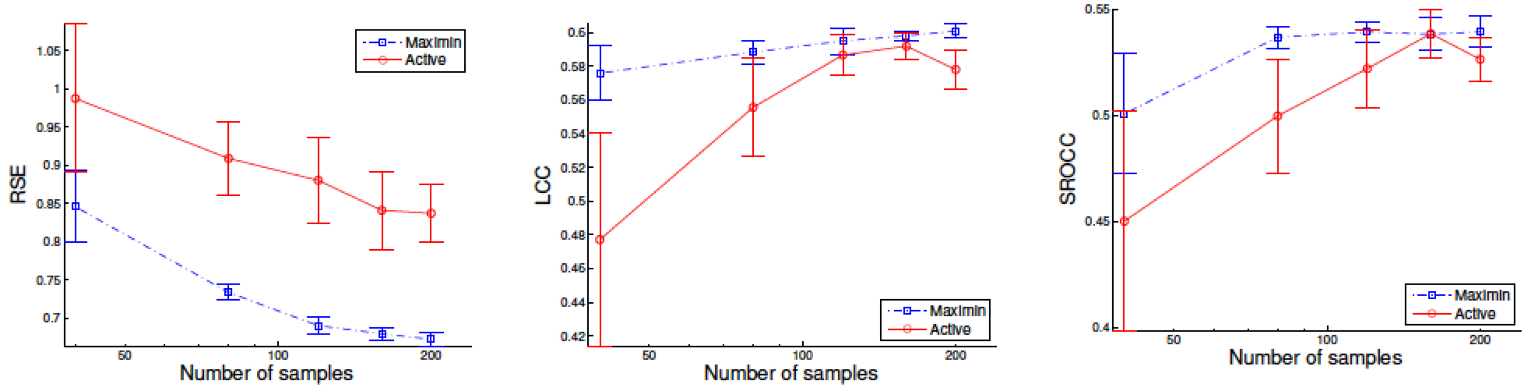
(b) Linear correlation coefficient.



(c) Spearman rank-order correlation coefficient.

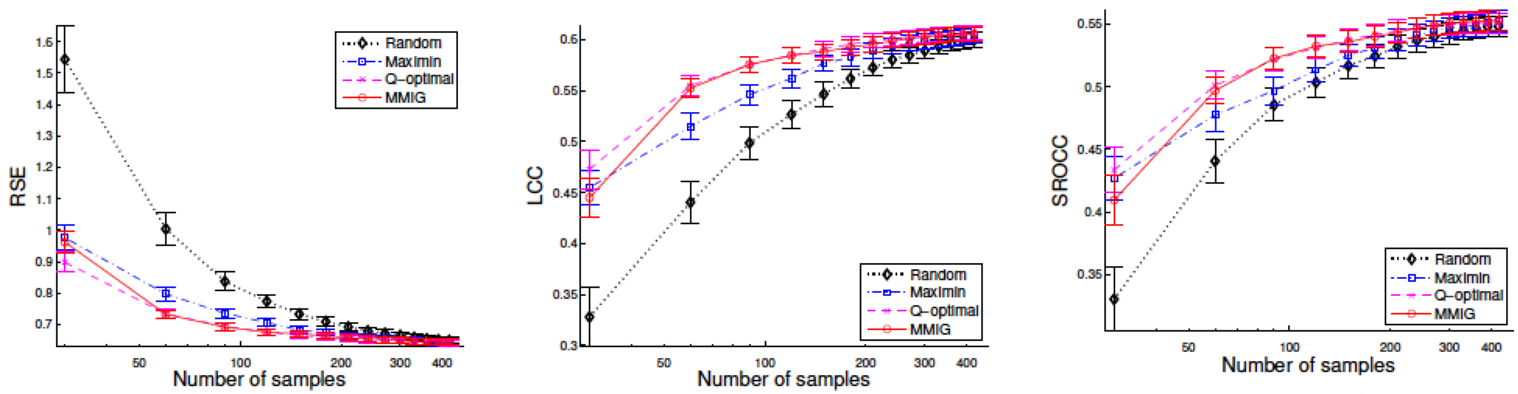
Active vs. Maximin Sampling

- Maximin > Q-optimal
  - Contradicting to the findings in offline setting



(a) Relative squared error. (b) Linear correlation coefficient. (c) Spearman rank-order correlation coefficient.

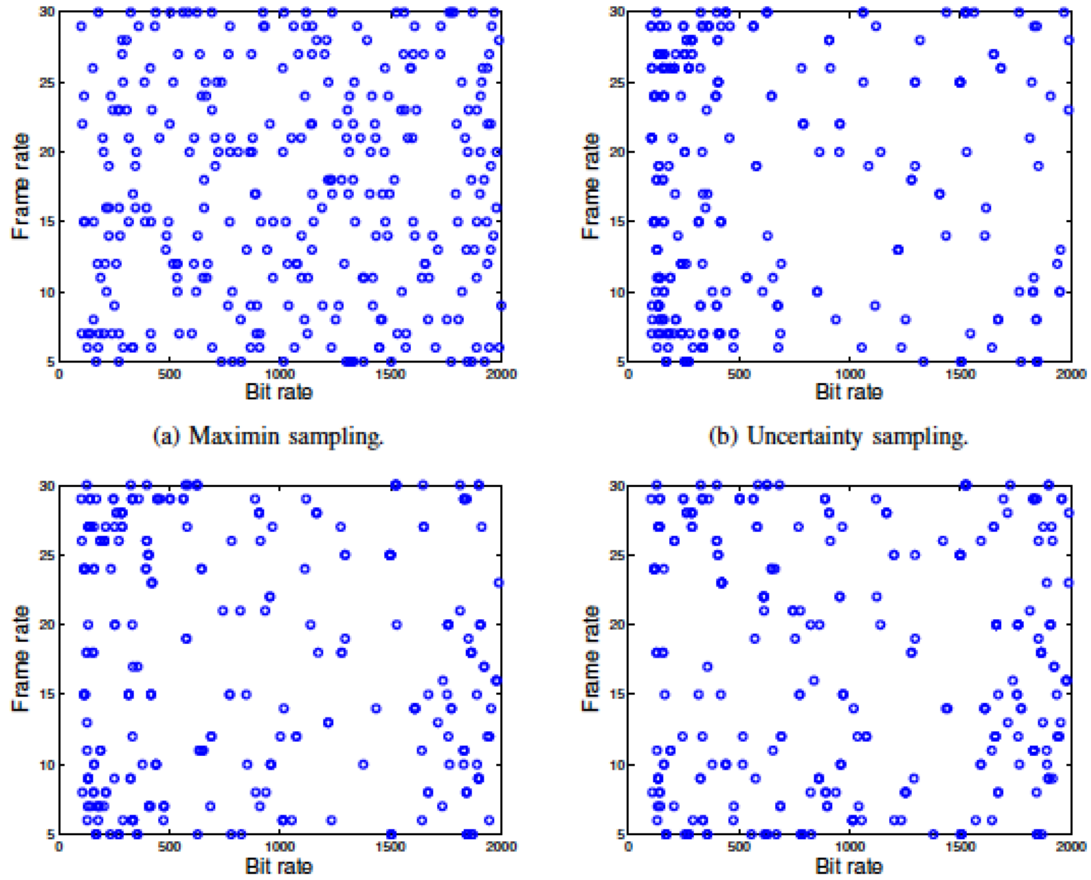
  - Repeat offline experiment: Q-optimal > Maximin



(a) Relative squared error. (b) Linear correlation coefficient. (c) Spearman rank-order correlation coefficient.

Difficulty I (Habituation Effect)

- Subjects tend to give higher scores than usual if they just saw a clip with very bad quality



(a) Maximin sampling.

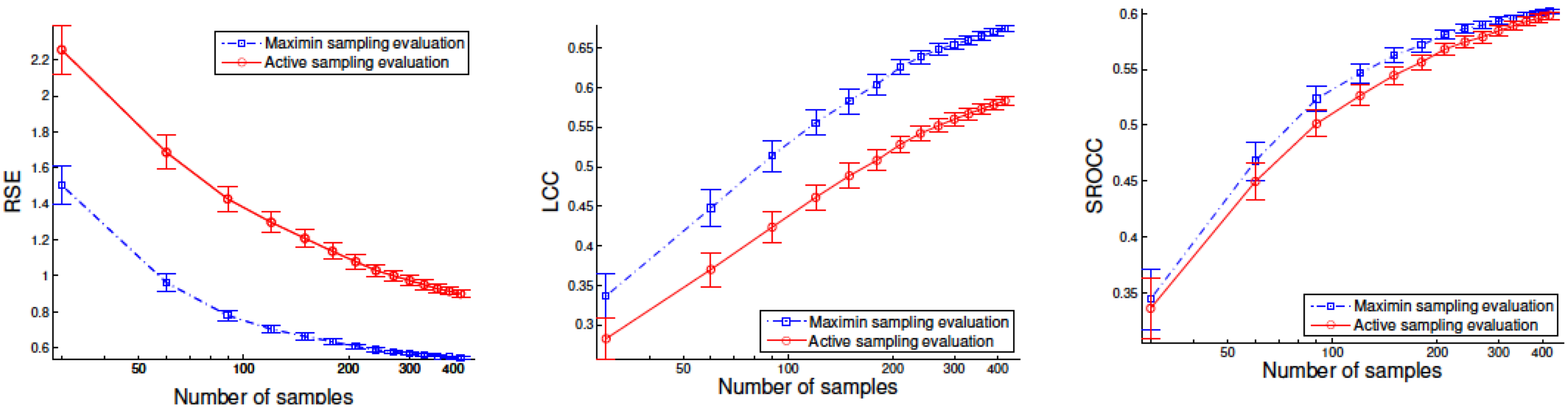
(b) Uncertainty sampling.

(c) Q-optimal sampling.

(d) MMIG sampling.

Difficulty I (Habituation Effect)

- MIQX model estimated using data from random sampling can predict scores from maximin sampling much better than it can predict scores from active sampling



(a) Relative squared error.

(b) Linear correlation coefficient.

(c) Spearman rank-order correlation coefficient.

Difficulty II (Individual Differences)

- Each subject has different standards for their judgement
- Active sampling has largest performance differences -> Active sampling might try to fit the QoE model of the current subject instead of fitting the average QoE model of the crowd

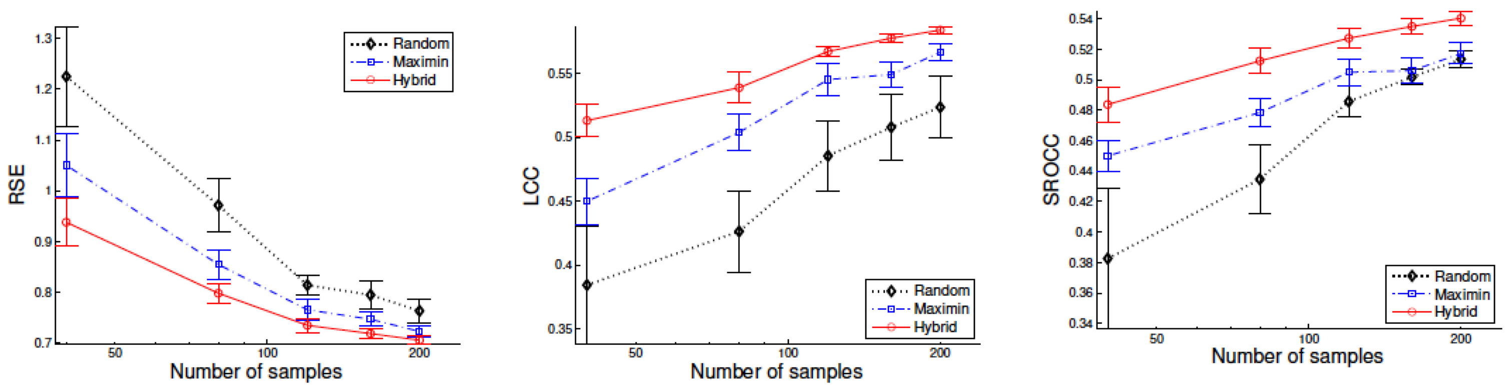
Method	Source of test data	RSE	LCC	SROCC
Random Sampling	The same trial	0.734 ± 0.014	0.567 ± 0.008	0.497 ± 0.008
	Different trials	0.839 ± 0.008	0.558 ± 0.003	0.489 ± 0.004
Active Sampling	The same trial	0.643 ± 0.011	0.637 ± 0.007	0.605 ± 0.007
	Different trials	0.742 ± 0.006	0.597 ± 0.003	0.565 ±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

Double Stimulus

- Reference video vs. compressed video
- Random, maximin, and hybrid (maximin+MMIG):

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

- 10 trials (200 samples labeled by 5 subjects)



(a) Relative squared error. (b) Linear correlation coefficient. (c) Spearman rank-order correlation coefficient.

Limitations

- Active learning still provides some bias in long-run

Training \ Testing	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

- The considered bitrate range
  - relative low compared with popular online video services, e.g., YouTube
  - The online workers might not have the adequate skills or hardware to identify the subtle difference among high-quality videos, e.g., 1 vs. 2 Mbps
  - Need to cover Larger interval

Conclusion

- Appropriate sampling methods are required to cope with the large parameter space
- Considering
  - Sampling strategies: random, maximin, active (uncertainty, q-optimal, and MMIG)
  - Models: linear regression, kernel regression, random forest, and MIQX
  - Testing methods: single stimulus and double stimulus

Issue

- Active learning may perform worse than passive learning due to habitual effect and individual differences
  - Active sampling+space-filling sampling
  - Take previous QoE scores into account
  - Model user diversity
  - Provide additional training for subjects
  - Filter unreliable subjects