Image Quality Assessment – Subjective Methods and Applications

Yuming Fang (方玉明)

School of Information Technology
Jiangxi University of Finance and Economics, Nanchang, China
江西财经大学 信息管理学院
Part I: Subjective Methods

• Introduction about image quality assessment

• Subjective datasets
The Quality of Multimedia Content

• Audition (hearing)

• Vision (seeing)

• Touch (taction)

• Olfaction (smell)

• Gustation (tasting)
What is Image Quality Assessment?
Synthetic and Authentic Distortions

Synthetic Distortions: Simulated by Pristine Image

Pristine image  |  BLUR: level 4  |  JPEG: level 4  |  JP2K: level 4

Realistic Distortion: Captured from Mobiles Devices

Smartphone Photography  |  Under-exposure  |  Motion blurring  |  Mixture distortions
Visual Quality Assessment

Subjective Quality Assessment
Getting human quality evaluation

Objective Quality Assessment
Output of a computational model

Observers

Environment

Methodology

Input visual data

Objective model

Prediction Results

Ground truth for training and/or validation
Visual Quality Assessment: Taxonomy

Visual Quality Assessment

• Subjective quality assessment
  – Reliable and accurate quality prediction of visual content
  – Time-consuming, laborious and expensive
  – Not applicable in practical applications

• Objective quality assessment
  – Predict perceived visual quality automatically
  – Applicable in practical applications
Subjective Image Quality Assessment

• Absolute category rating (ACR)
  – Single stimulus method
  – Test images are presented one at a time without reference information
  – Voting time: less or equal to 10 seconds depending on the voting method
  – Presentation time: 10 seconds depending on the test image content
  – Five-level or nine-level scale overall rating

• Absolute category rating with hidden reference (ACR-HR)
  – The only difference from the ACR method: a reference version of each test image must be included as the test stimulus, which is termed as a hidden reference condition

- 5 - Excellent
- 4 - Good
- 3 - Fair
- 2 - Poor
- 1 - Bad
Subjective Image Quality Assessment

- Degradation category rating (DCR)
  - Double stimulus method
  - Test images are presented in pairs: one is reference image, while the other is distorted image
  - Voting time: less or equal 10 seconds depending on voting method
  - Presentation time: 10 seconds depending on the image content
  - Five-level scale overall rating

<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Excellent</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
</tr>
<tr>
<td>2</td>
<td>Poor</td>
</tr>
<tr>
<td>1</td>
<td>Bad</td>
</tr>
</tbody>
</table>
Subjective Image Quality Assessment

• Pair comparison (PC)
  – Double stimulus method
  – Two test images from two different systems are presented in pair from the same reference image
  – Participants are asked to provide the judgment on which one is preferred in the test pair
  – All possible pairs are compared
    • N stimuli -> N(n-1)/2 pairs

- Which one do you prefer?
Subjective Image Quality Assessment

- Paired comparison model: converting paired comparison data to scale values

Pair comparison matrix

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>-</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>A2</td>
<td>6</td>
<td>-</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>A3</td>
<td>7</td>
<td>6</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>A4</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>-</td>
</tr>
</tbody>
</table>
Perceptual Visual Processing

Applications
- Gaming
- Video Conferencing
- Authentication

Technologies
- Perception-based Processing (denoising, enhancement, restoration, coding, transmission, etc.)
- Basic Computational Modules (CSF/VA/JND/……)
- Visual Quality Assessment (VQA)

Theories
- Visual Attention
- Visual Masking
- JND
- Eye Movement
- Contrast Sensitivity Function
- Perceptual Properties & Theories
- IPTV
- QoS Monitoring
Visual Quality Assessment: A Communication Framework

Acquisition
Prior knowledge about distortion: motion blur, noise, contrast...

Processing, Storage and Transmission
Prior knowledge about distortion: compression, packet loss, content loss...

Reconstruction and Display
Prior knowledge about the receiver and the task

Perception and Understanding

Visual Quality Measurement
LIVE Dataset

Some Reference Images in LIVE

- Distortion types: 5 (fast fading, Gaussian blur, JP2K, JPEG, white noise)

CSIQ Dataset

Some Reference Images in CSIQ

- Distortion types: 6 (JPEG, JP2K, Gaussian blur, white noise, contrast change, pink noise)

TID2013 Dataset

Some Reference Images in TID2013

- Reference images: 25.
- Distorted images: 3000.
- Distortion types: 24 (fast fading, Gaussian blur, JP2K, JPEG, white noise, etc.)

Some Reference Images in KADID-10K

- Distortion types: 25 (Gaussian blur, JP2K, JPEG, white noise, motion blur, etc.)

Waterloo Exploration Dataset

Some Reference Images in Waterloo Exploration

- Distortion types: 4 (Gaussian blur, JP2K, JPEG, White noise.)

Summary

- **Shortcoming**: we are often faced with the realistic distortions in real world, rather than the synthesized distortions.

- **Advantage**: it is very convenient/easy to build a large-scale database with diverse content, acting as the new independent test bed for IQA models or providing sufficient samples.
LIVE Challenge Dataset-Authentic Distortion

Some Samples in LIVE Challenge

- Distorted images: 1162.
- Distortion types: Complex.

KonIQ-10K Dataset-Authentic Distortion

Some Samples in KonIQ-10K

- Distorted images: 10073.
- Distortion types: Complex.

Smartphone Photography

Fast development of smartphone photography technologies:
• Hardware: Dual-camera system, wide-angle lens
• Software: HDR, portrait, panorama
We introduce a new image database, consisting of **11,125** pictures taken by 66 smartphones with 11 manufacturers.
We conduct so far the most comprehensive study of perceptual quality assessment of smartphone photography, including **image quality, image attributes** (brightness, colorfulness, contrast, noisiness, and sharpness), and **scene category labels** (animal, cityscape, human, indoor scene, landscape, night scene, plant, still life, and others).

Subjective user study for image quality and image attributes

- **First stage**: 1,125 images were rated by 104 sessions, and each session rated 80 images.
- **Second stage**: More than 600 subjects were invited to involved in the experiment, and each session rated 80 images (70 random selected from 10,000 images + 5 duplicated images + 5 images from first stage).
Subjective Experiments

Exchangeable image file format (EXIF) tags: 1) focal length, 2) f-number (inversely proportional to aperture size), 3) exposure time, 4) ISO (light sensitivity of sensor), 5) brightness value (brightness of focus point in scene), 6) flash (flash fired or not), 7) time (when image was recorded).

The EXIF data carry useful information about the scene being captured and the camera settings, which may help to predict the quality of smartphone photography.
Subjective Data Analysis

- Consistency across duplicated images: correlation between MOSs of duplicated images at two stages.
- Consistency across sub-groups: correlation between MOSs from two sub-groups of participants.
- Consistency across subjects: correlation between MOSs of individual participant and all participants.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Image attribute scores from subjects</th>
<th>Image attribute scores by MT-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brightness</td>
<td>0.784</td>
<td>0.704</td>
</tr>
<tr>
<td>Colorfulness</td>
<td>0.844</td>
<td>0.760</td>
</tr>
<tr>
<td>Contrast</td>
<td>0.874</td>
<td>0.786</td>
</tr>
<tr>
<td>Noisiness</td>
<td>0.893</td>
<td>0.832</td>
</tr>
<tr>
<td>Sharpness</td>
<td>0.958</td>
<td>0.904</td>
</tr>
</tbody>
</table>

The histogram of MOSs of the images in our database.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>SRCC</th>
<th>PLCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistency across duplicated images</td>
<td>0.893</td>
<td>0.903</td>
</tr>
<tr>
<td>Consistency across sub-groups</td>
<td>0.923</td>
<td>0.930</td>
</tr>
<tr>
<td>Consistency across subjects</td>
<td>0.841</td>
<td>0.865</td>
</tr>
</tbody>
</table>

- First column: SRCC values between MOSs and attribute scores from participants.
- Second column: SRCC values between MOSs and predicted attribute scores by MT-A.
Subjective Data Analysis

The top five and bottom five smartphone cameras based on image quality.

The MOS distribution of images in five quality levels for each scene category. Bad: MOS $\in [0, 19]$, Poor: MOS $\in [20, 39]$, Fair: MOS $\in [40, 59]$, Good: MOS $\in [60, 79]$, and Excellent: MOS $\in [80, 100]$.
Objective Quality Models

\[ \mathcal{L}_1(\mathcal{W}_B) = \|q - \hat{q}\|_1 = \sum_{i=1}^{m} |q^i - \hat{q}^i| \]

\( \mathcal{W}_B \) indicates the parameters in the baseline model (BL), and \( q^i \) is the MOS of \( i \)-th image.

\[ \mathcal{L}_2(\mathcal{W}_B, \mathcal{W}_A) = \beta_1 \|q - \hat{q}\|_1 + \frac{\beta_2}{5} \|r_j - \hat{r}_j\|_1 \]

\( \mathcal{W}_A \) indicates the parameters in the multi-task attribute (MT-A) model, and \( r^i \) is the attribute scores of \( i \)-th image. \( \beta_1 + \beta_2 = 1 \).

\[ \mathcal{L}_3(\mathcal{W}_B, \mathcal{W}_E) = \alpha_1 \|q - \hat{g}\|_1 + \alpha_2 \|q - \hat{q}\|_1 \]

\( \mathcal{W}_E \) indicates the parameters in the multi-task EXIF (MT-E) model. \( \alpha_1 = \alpha_2 = 0.5 \).
Objective Quality Models

Objective quality model of ResNet-50:

\[ L_1(W_B) = \| q - \hat{q} \|_1 = \sum_{i=1}^{m} |q^i - \hat{q}^i| \]

\[ L_4(W_S) = - \sum_{i,j} p_j^{(i)} \hat{p}_j^{(i)} \]

\[ L_5(W_B, W_S) = \frac{L_1(W_B)}{\sigma_1} + \frac{L_4(W_S)}{\sigma_2} + m \log \sigma_1 + \frac{m}{2} \log \sigma_2 \]

\( W_S \) indicates the parameters in the multi-task scene semantic (MT-S) model, and \( p^i \) is the scene category label of \( i-th \) image. \( \sigma_1 \) and \( \alpha_2 \) are two jointly learning parameters, which help to balance the image quality regression task and image classification task.
# Performance Evaluation

## Average PLCC and SRCC Results of Our Methods against Seven BIQA Models on SPAQ

<table>
<thead>
<tr>
<th>Model</th>
<th>QAC</th>
<th>DIIVINE</th>
<th>CORNIA</th>
<th>ILNIQE</th>
<th>BRISQUE</th>
<th>FRIQUEE</th>
<th>DB-CNN</th>
<th>BL</th>
<th>MT-A</th>
<th>MT-E</th>
<th>MT-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLCC</td>
<td>0.092</td>
<td>0.599</td>
<td>0.709</td>
<td>0.713</td>
<td>0.809</td>
<td>0.819</td>
<td>0.911</td>
<td>0.908</td>
<td>0.916</td>
<td>0.926</td>
<td>0.917</td>
</tr>
<tr>
<td>SRCC</td>
<td>0.497</td>
<td>0.600</td>
<td>0.725</td>
<td>0.721</td>
<td>0.817</td>
<td>0.830</td>
<td>0.915</td>
<td>0.909</td>
<td>0.916</td>
<td>0.932</td>
<td>0.921</td>
</tr>
</tbody>
</table>

## Average PLCC and SRCC Results of the Proposed BL Model in a Cross-Database Setting

<table>
<thead>
<tr>
<th>Training</th>
<th>SPAQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Synthetic database</td>
</tr>
<tr>
<td></td>
<td>LIVE</td>
</tr>
<tr>
<td>PLCC</td>
<td>0.608</td>
</tr>
<tr>
<td>SRCC</td>
<td>0.560</td>
</tr>
</tbody>
</table>
Summary

- BIQA models designed for synthetic distortions (e.g., QAC and DIIVINE) generally do not work well for realistic camera distortions.
- Verified on the LIVE Challenge Database, FRIQUEE delivers superior performance on the proposed database.
- DBCNN outperforms all BIQA approaches, including the proposed BL based on ResNet-50, which suggests that DNNs successfully learn hierarchical sensitive features to realistic distortions.
- Image attributes (MT-A) positively impacts the accuracy of quality prediction.
- MT-E achieves a significant improvement compared with BL. This emphasizes the importance of EXIF data to quality prediction of smartphone captured images, which, however, has not been paid much attention by the IQA community.
- MT-S is able to exploit semantic information to boost the quality prediction performance. These insightful findings inspire further research on how to extract semantic information.

Database & Models:
https://github.com/h4nwei/SPAQ

Conventional subjective testing requires manually pre-selecting a small set of visual examples, which may suffer from three sources of biases:

- *Sampling bias* due to the extremely sparse distribution of the selected samples in the image space;
- *Algorithm bias* due to potential overfitting the selected samples;
- *Subjective bias* due to further potential cherry-picking test results.

Target: debiased subjective assessment.
Debiased Subjective Assessment of Real-World Image Enhancement

Overview of our debiased subjective assessment in the context of single image dehazing
(a) A large set of hazy images. (b) Top-$K$ hazy images selected from (a) to best discriminate between Shao20 and FFANet by optimizing Eq. (1). (c) Pairs of dehazed images corresponding to representative hazy images in (b).

\[
\hat{x}^{(k)} = \arg \max_{x \in \chi \setminus S} D_1(f_1(x), f_2(x)) + \lambda_1 D_2(x, S) \quad (1)
\]

Global ranking results of (a) single image dehazing, (b) single image super-resolution, and (c) low-light image enhancement.