Advanced Graphics & Image Processing

Assessing Image Quality

Rafał Mantiuk
Computer Laboratory, University of Cambridge
The purpose of image quality assessment

- To compare algorithms in terms of image or video quality

Rate-Distortion (RD) curves
The purpose of image quality assessment

- To optimize application parameters – e.g. resolution and bit-rate
The purpose of image quality assessment

- To provide evidence of improvement over the state-of-the-art
Other application domains

- Recommendation systems
  - Which movie to watch? (Netflix)
  - Which product to buy? (Amazon)

- Product acceptance / rating
  - Food
  - Clothing
  - Consumer electronics, …

- Similar techniques used for
  - Ranking of the players/gamers to match their skills in the game (TrueSkill on Xbox)
Subjective image/video quality assessment methods

Subjective quality assessment

- Ranking
  - Ordinal scaling
    - Rank order method
    - Pair-wise comparisons
    - ... (other methods)

- Rating
  - Direct interval scaling
    - Single stimulus with hidden reference
    - Double stimulus
    - ... (other methods)
Rating: Single stimulus + hidden reference

- With a hidden reference
- Task: **Rate** the quality of the image
- The categorical variables (excellent, good, ...) are converted into scores 1-5
- Then those are averaged across all observers to get Mean-Opinion-Scores (MOS)
- To remove the effect of reference content, we often calculate DMOS:

\[ Q_{DMOS} = Q_{MOS}^{reference} - Q_{MOS}^{test} \]
Rating: Double stimulus

- Task: Rate the quality of the first and the second image
- The second image is typically the reference
- Potentially better accuracy of DMOS
- But takes more time
  - The reference shown after each test image
Pair-wise comparison method

- Example: video quality
- Task: Select the video sequence that has a higher quality
Comparison matrix

- Results of pairwise comparisons can be stored in a comparison matrix

\[
C = \begin{bmatrix}
0 & 3 & 1 \\
3 & 0 & 2 \\
5 & 4 & 0
\end{bmatrix}
\]

- In this example: 3 compared conditions: C1, C2, C3
- \( C_{ij} = n \) means that condition C_\text{i} was preferred over C_\text{j} \( n \) times
Full and reduced designs

- **Full design**
  - Compare all pairs of conditions
  - This requires \( \binom{n}{2} = \frac{n(n-1)}{2} \) comparisons for \( n \) conditions
  - Tedious if \( n \) is large

- **Reduced design**
  - We assume transitivity
    - If \( C_1 > C_2 \) and \( C_2 > C_3 \) then \( C_1 > C_3 \)
      - no need to do all comparisons
  - There are numerous “block designs” (before computers)
  - But the task is also a sorting problem
    - The number comparison can be reduced to \( n \log(n) \) for a “human quick-sort”
  - And many others: Swiss chess system, active sampling ...
Pairwise comparisons vs. rating (e.g. single stimulus)

- The method of pairwise comparisons is **fast**
  - More comparisons, but
  - It takes less time to achieve the same sensitivity as for direct rating methods
- Has a higher sensitivity
  - Less “external” variance between and within observers
- Provides a unified quality scale
  - The scale (of JOD/JND) is transferrable between experiments
- Simple procedure
  - Training is much easier
  - Less affected by learning effects
- Especially suitable for non-expert participants
  - E.g. Crowdsourcing experiments
Time-efficiency

The results show how long (on average) it took participants to complete the experiment.
Active sampling can make the experiments even faster

- **Active sampling**
  - For each trial, select a pair of conditions that maximizes the information gain
  - Information gain is the DK-divergence between the prior and posterior distributions

Practical significance - scaling

- Scaling: to map user judgments into meaningful interval scale
- Typically that scale is in just-noticeable-difference units
  - The difference of 1 JND means that 75% of observers would choose one condition over another
- Useful to show “practical” significance
Scaling pairwise comparison data

- Given a matrix of comparisons, for example

\[
C = \begin{bmatrix}
0 & 3 & 0 \\
27 & 0 & 7 \\
30 & 23 & 0
\end{bmatrix}
\]

- Infer the quality scores for all compared conditions
  - Using Maximum Likelihood Estimation (MLE)

- We start from an observer model, then link it to the observations
Thurstone (observer) model - Case V

- **Two assumptions:**
  - Quality scores for a given condition are normally distributed across the population.
  - The variance of that distribution is the same for each condition and the judgements are independent.
From the observer model to probabilities

- Given the observer model for two conditions:
  \[ r_i = N(q_i, \sigma^2) \quad r_j = N(q_j, \sigma^2) \]

- The difference between two quality scores is:
  \[ r_i - r_j = N(q_i - q_j, 2\sigma^2) \]

- Then, the probability of the judgment is explained by the cumulative normal distribution
  \[
P(r_i > r_j) = P(r_i - r_j > 0) = \Phi \left( \frac{q_i - q_j}{\sigma_{ij}} \right)
  = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q_i - q_j} e^{-\frac{x^2}{2\sigma_{ij}^2}} dx.
  \]
  where \( \sigma_{ij} = \sqrt{2}\sigma \)
Given that $k$ out of $n$ observers selected A over B, what is the probability distribution of selecting A over B

\[ P(r_i > r_j | n, k) = \binom{n}{k} p^k (1 - p)^{n-k} \]
Maximum Likelihood Estimation

- Given our observations (comparison matrix) what is the likelihood of the quality values $q_i$:

$$L(\hat{q}_i - \hat{q}_j | c_{ij}, n_{ij}) = \frac{n_{ij}}{c_{ij}} P(r_i > r_j)^{c_{ij}} (1 - P(r_i > r_j))^{n_{ij} - c_{ij}}$$

$$= \frac{n_{ij}}{c_{ij}} \Phi \left( \frac{\hat{q}_i - \hat{q}_j}{\sigma_{ij}} \right)^{c_{ij}} \left( 1 - \Phi \left( \frac{\hat{q}_i - \hat{q}_j}{\sigma_{ij}} \right) \right)^{n_{ij} - c_{ij}}$$

- where $n_{ij} = c_{ij} + c_{ji}$

- To estimate the values of $q_i$, we maximize:

$$\arg \max_{\hat{q}_2, \ldots, \hat{q}_n} \prod_{i, j \in \Omega} L(\hat{q}_i - \hat{q}_j | c_{ij}, n_{ij})$$
JND/JOD = 1

- Just Noticeable Differences
- Just Objectionable Differences
- We want $q_i - q_j = 1$ when 75% of observers prefer condition “i” over “j”

- This happens when $\sigma_{ij} = 1.4826$
- This is arbitrary selected scaling, made for easier interpretation of the results
JND vs JOD

- Just Noticeable Differences
- Just Objectionable Differences

- JND – is one visually different from another
- JOD – is the quality of one different from the quality of another (relative to the reference)
Practicalities of MLE scaling

- At least 15-20 comparisons per each pair are needed to obtain stable results (prior helps)
Forced choice vs. comparison with ties

- Giving a “tie” option is usually a bad idea
- Scaling the results with ties requires a more complex observer model with more parameters to estimate
Objective (image/video) quality metrics
Types of objective (image/video) quality metrics

**Full Reference (FR) metrics**
- Test image
- Reference image
- Full-reference quality metric
- Quality score
- (optional) Distortion map

**No Reference (NR) metrics**
- Test image
- No-reference quality metric
- Quality score

**Reduced Reference (RR) metrics**
- Test image
- Reference image
- Reduced-reference quality metric
- Quality score
- Image statistics
Main use cases of objective quality metrics

(I) Evaluation

Which method is the best?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Set5</td>
<td>×2</td>
<td>33.66 / 0.9299</td>
<td>36.54 / 0.9544</td>
<td>36.66 / 0.9542</td>
<td>37.53 / 0.9587</td>
</tr>
<tr>
<td></td>
<td>×3</td>
<td>30.39 / 0.8842</td>
<td>32.58 / 0.9088</td>
<td>32.75 / 0.9099</td>
<td>33.66 / 0.9213</td>
</tr>
<tr>
<td></td>
<td>×4</td>
<td>28.42 / 0.8104</td>
<td>30.28 / 0.8603</td>
<td>30.48 / 0.8628</td>
<td>31.35 / 0.8838</td>
</tr>
<tr>
<td>Set14</td>
<td>×2</td>
<td>30.24 / 0.8688</td>
<td>32.28 / 0.9056</td>
<td>32.42 / 0.9063</td>
<td>33.03 / 0.9124</td>
</tr>
<tr>
<td></td>
<td>×3</td>
<td>27.55 / 0.7742</td>
<td>29.13 / 0.8188</td>
<td>29.28 / 0.8209</td>
<td>29.77 / 0.8314</td>
</tr>
<tr>
<td></td>
<td>×4</td>
<td>26.00 / 0.7027</td>
<td>27.32 / 0.7491</td>
<td>27.49 / 0.7503</td>
<td>28.01 / 0.7674</td>
</tr>
<tr>
<td>B100</td>
<td>×2</td>
<td>29.56 / 0.8431</td>
<td>31.21 / 0.8863</td>
<td>31.36 / 0.8879</td>
<td>31.90 / 0.8960</td>
</tr>
<tr>
<td></td>
<td>×3</td>
<td>27.21 / 0.7385</td>
<td>28.29 / 0.7855</td>
<td>28.41 / 0.7863</td>
<td>28.82 / 0.7976</td>
</tr>
<tr>
<td></td>
<td>×4</td>
<td>25.96 / 0.6675</td>
<td>26.82 / 0.7087</td>
<td>26.90 / 0.7101</td>
<td>27.29 / 0.7251</td>
</tr>
<tr>
<td>Urban100</td>
<td>×2</td>
<td>26.88 / 0.8403</td>
<td>29.20 / 0.8938</td>
<td>29.50 / 0.8946</td>
<td>30.76 / 0.9140</td>
</tr>
<tr>
<td></td>
<td>×3</td>
<td>24.46 / 0.7349</td>
<td>26.03 / 0.7973</td>
<td>26.24 / 0.7989</td>
<td>27.14 / 0.8279</td>
</tr>
<tr>
<td></td>
<td>×4</td>
<td>23.14 / 0.6577</td>
<td>24.32 / 0.7183</td>
<td>24.52 / 0.7221</td>
<td>25.18 / 0.7524</td>
</tr>
</tbody>
</table>

Aims:
- To demonstrate the difference in quality
- To replace subjective experiments

(II) Optimization

What are the best parameter values?

Aims:
- To replace manual parameter tweaking
- Especially in multi-dimensional problems
Pixel-wise quality metrics

- **Root Mean Square Error (RMSE)**
  \[ E_{RMSE} = \sqrt{\frac{1}{w \cdot h} \sum_{x,y} (t(x,y) - r(x,y))^2} \]

- **Peak Signal to Noise Ratio (PSNR)**
  \[ E_{PSNR} = 20 \frac{I_{peak}}{E_{RMSE}} \text{ [dB]} \]
  - \( I_{peak} \) - the peak pixel value (e.g. 255 or 1)
  - If the error is normally distributed and its mean is 0, \( E_{RMSE} \) is the standard deviation of the distortion (noise)
The shortcomings of pixel-wise metrics

Reference

JPEG-encoded
PSNR=24.7

Blur
PSNR=24.8

Noise
PSNR=24.8

Rotation (1.3 deg)
PSNR=23.4

[Examples from: 10.1109/TIP.2008.926161]
Texture quality metrics

Test image

Extract (local) image statistics (e.g. mean, var)

Reference image

Extract (local) image statistics (e.g. mean, var)

Pooling

Quality score

≠ per pixel

≈ appearance

30
Structural Similarity Index (SSIM)

- Split test and reference images into $11 \times 11$ px overlapping patches
- For each patch, calculate mean $\mu_T, \mu_R$, std $\sigma_T, \sigma_R$ and covariance $\sigma_{TR}$ of each patch, weighted by a Gaussian window
- Calculate three terms (per patch)
  - “Luminance”: $l_x = \frac{2\mu_T \mu_R + C_0}{\mu_T^2 + \mu_R^2 + C_0}$
  - Contrast: $c_x = \frac{2\sigma_T \sigma_R + C_1}{\sigma_T^2 + \sigma_R^2 + C_1}$
  - Structure: $s_x = \frac{\sigma_{TR} + C_2}{\sigma_T \sigma_R + C_2}$ (cross-correlation)
- Multiply them together: $q_x = l_x \cdot c_x \cdot s_x$
- And pool: $q_{SSIM} = \frac{1}{N} \sum x q_x$
Learned Perceptual Image Patch Similarity (LPIPS)

- Use a pre-trained CNN as a feature extractor

---

Diagram:
- Test image $x$
- Reference image $x_0$
- Learned weights $w$
- Feature differences
- Multiply L2 norm Spatial Average
- Average $d_0$
- Predicted quality
Metrics and viewing conditions

- Majority of image/video metrics disregard viewing conditions
  - Display size
  - Display resolution
  - Viewing distance
  - Display peak luminance
  - Colour gamut
- PSNR, SSIM, LPIPS operate on 0-255 pixel values
  - Cannot handle HDR images/video
- To account for the viewing conditions, we need metrics based on psychophysical models
  - Known as visual difference predictors (VDPs)
Perceptual metrics (Visual Difference Predictors)

"standard_4k": {
    "resolution": [3840, 2160],
    "viewing_distance_meters": 0.7472,
    "diagonal_size_inches": 30,
    "max_luminance": 200,
    "contrast": 1000,
    "E_ambient": 250,
}
Perceptual metrics (Visual Difference Predictors)
Perceptual metrics (Visual Difference Predictors)
Perceptual metrics (Visual Difference Predictors)
Perceptual metrics (Visual Difference Predictors)

Test contrast

Mask contrast

castleCSF
minimum
detectable
contrast
difference
Perceptual metrics (Visual Difference Predictors)

The quality is scaled in the units of Just Objectionable Differences [JOD]
1 JOD difference ≈ 50% increase in preference

Can express supra-threshold (well-visible) differences
Metric performance on band-limited noise

40  Violet – large difference; Orange – small difference
Metric performance on masking patterns

![Graphs showing metric performance on masking patterns]

- **Violet** – large difference;
- **Orange** – small difference

Contrast of the distortion (Gabor)
Contrast of the masker
References

- Scaling of pairwise comparison data
  - pwcmp - https://github.com/mantiuk/pwcmp
  - A practical guide and software for analysing pairwise comparison experiments - https://arxiv.org/abs/1712.03686

- Active sampling
  - ASAP - https://github.com/gfxdisp/asap

- SSIM
  - A Hitchhiker’s Guide to Structural Similarity - https://doi.org/10.1109/ACCESS.2021.3056504

- VDP metrics
  - HDR-VDP – https://hdrvdp.sourceforge.net/
  - FovVideoVDP - https://github.com/gfxdisp/FovVideoVDP
  - ColorVideoVDP - https://github.com/gfxdisp/ColorVideoVDP