JPEG AIC-3 Activity on fine-grained assessment of subjective quality of compressed images

Michela Testolina, Vlad Hosu, Mohsen Jenadeleh, Davi Lazzarotto, Dietmar Saupe, Touradj Ebrahimi
+ Shaolin Su, Oliver Wiedemann, Hui Men
JPEG AIC-3 Dataset

- 10 reference images, different resolutions and content
- Compression artifacts generated with JPEG, JPEG 2000, HEVC Intra, VVC Intra, and JPEG XL at multiple quality levels
- Visual quality range from high to nearly visually lossless
  - Selected through a subjective image quality assessment experiment
Subjective experiment

- A preliminary subset of distorted images was selected by visual inspection
  - Statistical analysis and interpolation to refine the initial selection and extract the final dataset
- Conducted in a crowdsourcing environment with expert viewers
- Minimum screen size 1920×1080, retina mode disabled
- Image cropping to a size of 945×880
Subjective experiment

- Protocol: variation of the pair comparison (PC) experiment
- Subjects were asked to select the stimulus presenting the highest visual quality between two options, displayed side-by-side.

![Sample A | Sample B](image)

Please select the image with the highest visual quality:
- Sample A
- Not sure
- Sample B

Next
**Statistical analysis**

- **JND** values were reconstructed from the collected subjective visual scores

- An analysis similar to [1] was adopted:
  - Standard reconstruction was applied by maximum likelihood estimation according to the *Thurstonian probabilistic* model (Case V)
  - Results were scaled to JND units
    - If two images are 1 JND unit apart, then the model predicts a **50% probability for the detection of the difference** by a random observer

---

Statistical analysis

- From the JND scores collected on the preliminary subset, the selection is refined targeting images in the visual quality range of interest.
- A parametric curve was fitted to the collected subjective quality scores:
  - Sum of a linear a logistic function

\[
f(x) = -a \left(1 - \frac{x}{100}\right) + \frac{100}{1 + e^{-100b\left(\frac{x}{100}-c\right)}} - 100
\]

- The selected minimum scale value is -2.5 JND.
- The scale interval [-2.5,0] was subdivided into 10 subintervals of equal 0.25 JND length.
Statistical analysis

Thurstone reconstructed score (in JND units)

Quality normalized to [0,100]
Fine-grained assessment of subjective quality of compressed images

Vlad Hosu¹, Mohsen Jenadeleh¹, Shaolin Su¹, Oliver Wiedemann, Hui Men, Dietmar Saupe, University of Konstanz

¹The first three authors contributed equally.
Our proposal: Boosted triplet comparison

Received July 31, 2021, accepted September 3, 2021, date of publication October 5, 2021, date of current version October 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3118295

Subjective Image Quality Assessment With Boosted Triplet Comparisons

HUI MEN®, HANHE LIN®, MOHSEN JENADELEH®, (Member, IEEE), AND DIETMAR SAUPE®

Department of Computer and Information Science, University of Konstanz, 78464 Konstanz, Germany

Corresponding author: Hui Men (hui.3.men@uni-konstanz.de)

This work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Project 251654672—TRR 161 (Project A05).
Reference and Distorted Image

Ref.

Orig. Dist.
Boosting (A)

\[ v' = v_{ref} + \alpha(v_{\text{dist}} - v_{ref}) \ (\alpha > 1) \]
Boosting (A+Z)

Ref.

Added Zoom (Z)
Boosting (A+Z+F)
Comparison of two compressed images flickering w.r.t. source image

Which image has a stronger flicker effect?

left  not sure  right
Impairment Scales for JPEG2000, A+F, 10 images and average
Application for JPEG AIC-3 dataset

- 10 source images
- 6 codecs
- 10+1 distortion levels (estimated at 0.25*k JND, k=0,...,10)
- 60 image sequences of 11 images each
Baseline triplet comparisons
Artefact amplification and flicker test

- Baseline triplets are \((i,0,k)\)
  - Two images at levels \(i\) and \(k\) are compared with the source (level 0)
- Same-codec and cross-codec comparisons
- Selection of triplet comparisons:
  - Per sequence of 11 images: All 110 triplets \((i,0,k)\) with \(i < k\) or \(k < i\).
  - This makes \(60 \times 110 = 6600\) same-codec triplets
  - Recommendation to include cross-codec comparisons (randomly choose codecs and levels) [E. Zerman, QoMEX 2019]: 1200 triplets
  - Random triplets \((10,0,0)\) and \((0,0,10)\) as trap questions: 780 triplets
  - Total number \(6600 + 1200 + 780 = 8580\) triplets

Crowdsourcing Campaign

- AMT platform (mturk)
- 110 study questions per HIT
  - 100 study questions, 10 trap questions
  - Each study question in both orientations: (i,0,k) and (k,0,i)
  - 8580 triplets / 110 = 78 HITs
- Deploy each HIT with 30 assignments
  - Collect 30*78*110 = 257400 responses
- Quality control
  - Require 98% acceptance rate in previous work of crowd workers
  - Minimum screen resolution of 1920x1080 pixels
- Timing
  - 5 + 3 seconds per triplet (no answer in 8 secs -> „skipped response“)
  - 30 minutes per assignment
View of a crowdworker at mturk
Accuracy and consistency: Definitions

- Accuracy := ratio of correct answers for all triplets of type \((0,0,10)\) and \((10,0,0)\)

- Consistency := ratio of consistent responses to the 50 triplet pairs \((i,0,k)\) and \((k,0,i)\)

Accuracy and consistency for 2265 assignments
Data filtering and outlier removal

- Filtering and outlier removal at assignment level (110 triplets each)
- Assignments will be included if all of following hold:
  - Number of skipped questions <= 10
  - Accuracy >= 0.7
  - Consistency >= 0.6
- Iterative outlier removal for the remaining assignments based on negative log-likelihood (NLL)
  - Get statistical data model by MLE of the minimum of the global NLL
  - Compute the NLL for all assignments (including outlier candidates)
  - Mark assignments outside the 90th percentile as outlier candidates
  - Repeat until convergence
Empirical probabilities from experiment

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5000</td>
<td>0.8000</td>
<td>0.9524</td>
<td>1.0000</td>
<td>0.9800</td>
<td>1.0000</td>
<td>0.9800</td>
<td>1.0000</td>
<td>0.9800</td>
<td>0.9783</td>
<td>1.0000</td>
</tr>
<tr>
<td>2</td>
<td>0.0750</td>
<td>0.5000</td>
<td>0.6944</td>
<td>0.9286</td>
<td>0.8889</td>
<td>0.9134</td>
<td>0.8580</td>
<td>0.9056</td>
<td>0.9500</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>3</td>
<td>0.0714</td>
<td>0.1667</td>
<td>0.5000</td>
<td>0.6875</td>
<td>0.8056</td>
<td>0.9500</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>4</td>
<td>0.0476</td>
<td>0.0476</td>
<td>0.1667</td>
<td>0.5000</td>
<td>0.5833</td>
<td>0.8235</td>
<td>0.8478</td>
<td>0.9600</td>
<td>0.9412</td>
<td>0.9737</td>
<td>1.0000</td>
</tr>
<tr>
<td>5</td>
<td>0.0227</td>
<td>0.0000</td>
<td>0.1944</td>
<td>0.5000</td>
<td>0.6304</td>
<td>0.8611</td>
<td>0.9063</td>
<td>0.9474</td>
<td>0.9348</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>6</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1304</td>
<td>0.5000</td>
<td>0.7045</td>
<td>0.8611</td>
<td>0.9538</td>
<td>0.7500</td>
<td>0.9545</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>7</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0455</td>
<td>0.5000</td>
<td>0.3864</td>
<td>0.5000</td>
<td>0.5714</td>
<td>0.5000</td>
<td>0.7174</td>
<td>0.9118</td>
<td>1.0000</td>
</tr>
<tr>
<td>8</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.5000</td>
<td>0.1053</td>
<td>0.1154</td>
<td>0.0526</td>
<td>0.2222</td>
<td>0.5000</td>
<td>0.4375</td>
<td>0.6667</td>
</tr>
<tr>
<td>9</td>
<td>0.0000</td>
<td>0.0208</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1042</td>
<td>0.0217</td>
<td>0.0294</td>
<td>0.2917</td>
<td>0.5000</td>
<td>0.6750</td>
<td>1.0000</td>
</tr>
<tr>
<td>10</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0435</td>
<td>0.0227</td>
<td>0.0600</td>
<td>0.1111</td>
<td>0.2250</td>
<td>0.5000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Example: Image 0, Codec 1

The empirical probabilities on the diagonal are not from the experiment. Stimuli were not compared with themselves. These values 0.5 are included only to help Matlab to create the heatmap correctly.
Perceived distortion vs distortion level
Perceived distortion vs. level: Summary

Mean over 10 sources

- AVIF
- HM
- JPEG
- JPEG2000
- JPEG-XL
- VVC
Perceived distortion vs bitrate
Ongoing work: Core experiment

- Crowdsourcing study 1: Triplet comparisons including also 2x zoom on crops
- Crowdsourcing study 2: Double Stimulus Boosted Quality Scale (DSBQS) protocol
  - Subject can toggle view between source and compressed image (twice per second)
  - Subject rates quality of compressed image on an interval scale
- Unified statistical model for
  - Data cleansing / outlier removal
  - Merging of the two datasets
End