Impact of Feedback on Crowdsourced Visual Quality Assessment with Paired Comparisons

Mohsen Jenadeleh, Alexander Heß, Simon Hviid Del Pin, Edwin Gamboa, Matthias Hirth, Dietmar Saupe

University of Konstanz, Germany
Norwegian University of Science and Technology, Norway
ScaleHub GmbH, Germany
Ilmenau University of Technology, Germany

June 18, 2024

Background

- **Crowdsourcing for image and video quality assessent**
- **Advantages of Crowdsourcing:**
  - Easy access to diverse populations
  - Scalability for large experiments
  - Time-efficient and cost-effective
- **Challenges in Crowdsourcing:**
  - Maintaining attention of crowdworkers
  - Ensuring high-quality, accurate data

Motivation

- **Need for Reliable Crowdsourcing Tasks:**
  - Ensure high-quality data collection from crowdworkers
  - Improve the efficiency and reliability of subjective image quality assessments
- **Immediate Feedback as Simple Gamification element:**
  - Provide real-time feedback on accuracy of the crowdworkers' responses
  - Increase engagement and motivation of crowdworkers

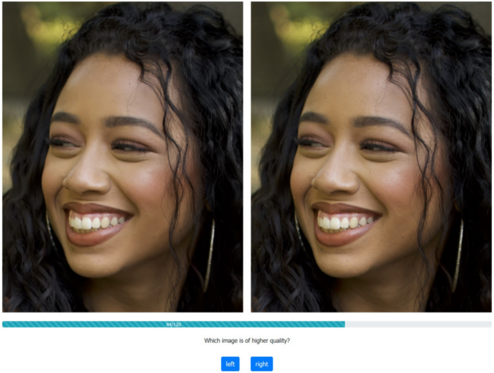
Hypotheses

Task: Just noticeable difference (JND) based image quality assessment

- **Hypothesis 1:** Feedback does not influence the JND estimation
- **Hypothesis 2:** Feedback increases the precision of JND assessment

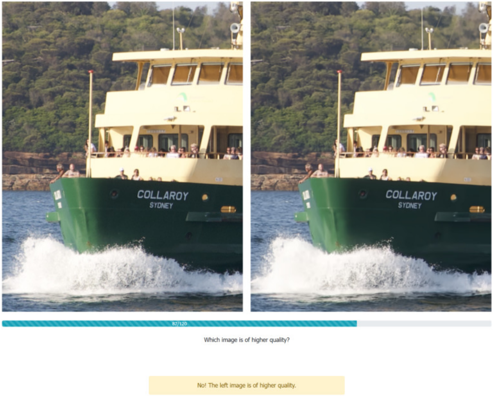
Experimental Setup

- With-in subject design experiment
- Test conditions:
  - 2AFC with feedback (AFCF)
  - 2AFC without feedback (AFCN)
- Order of conditions randomized for each subject
- Feedback provided in the form of text and audible beeps
- A cognitive load questionnaire followed the PCs for each condition.



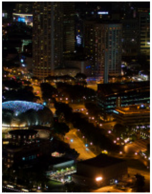
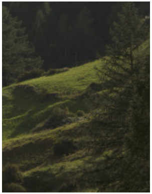
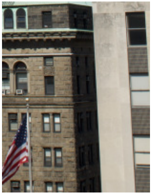
Feedback for correct response
Text message + low frequency audible beep

Feedback for wrong response
Text message + high frequency audible beep



Source and test images

- We used 5 source images from JPEG AIC dataset
- Compressed by two codecs, VVC Intra and JPEG XL at 10 distortion levels
- 5(sources) $\times$ 2(codecs) $\times$ 10(dist. levels) $=$ 100 PCs (questions)
- 20 trap questions (with strong distortion difference)
- 6 training questions

Crowdsourcing

- Conducted on Amazon Mechanical Turk (MTurk).
- Each worker completed 120 paired comparisons with and without feedback.
- Order of the questions are randomized for each participant
- Involved 200 crowdworkers (the responses of 149 workers accepted)

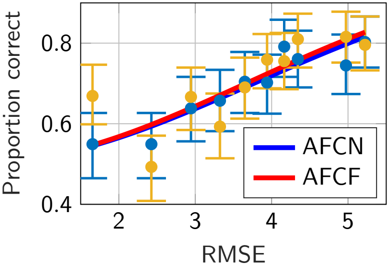
Statistical model for data analysis

- Weibull psychometric function (PSF) is fitted to the proportions of correct responses.

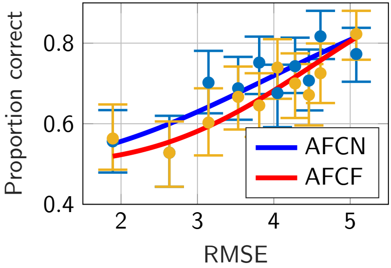
$$\psi(x;\lambda,k) = \frac{1}{2} + \frac{1}{2}\left(1 - e^{-(x/\lambda)^k}\right)$$

  - x: Independent variable (RMSE in a compressed image).
  - λ: Scale parameter.
  - k: Shape parameter.
  - JND reported at 75% proportion correct $\psi(x;\lambda,k) = 0.75$

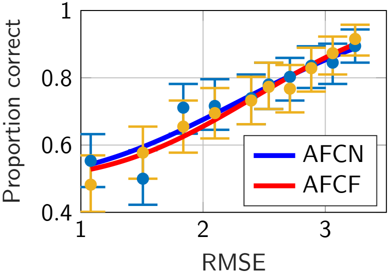
Results: Fitted PSF



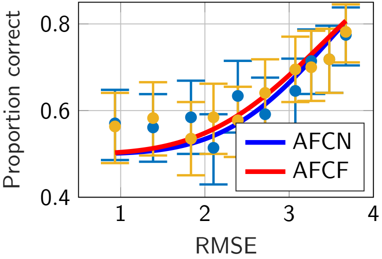
(a) S6-JPEG XL



(b) S10-JPEG XL



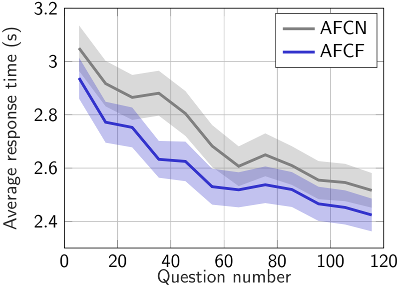
(c) S6-VVC



(d) S10-VVC

Results: Response time

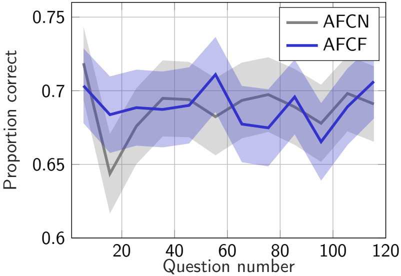
- **Average response time:**
  - With feedback (AFCF): 2.598 seconds
  - Without feedback (AFCN): 2.723 seconds
- **Significance:**
  - P-value: 0.0005
  - Feedback significantly reduced response time ("small" effect size, Cohen's d = 0.1).



Effect size	**Cohen's d**
small	0.20
medium	0.50
large	0.80
Very large	1.20

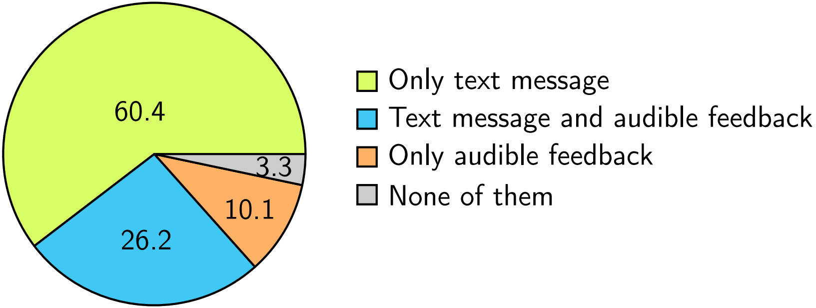
Results: Learning effect

- **Proportion of correct responses over time:**
  - No significant learning effect observed.
  - No significant difference between conditions with and without feedback.
- **Batch 2 Analysis:**
  - Lower proportion of correct responses.
  - Two-proportion Z-test: Z-score = 0.45, p-value = 0.65.



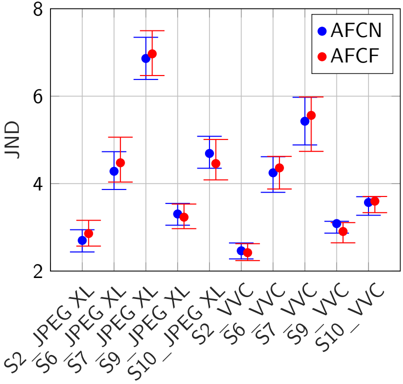
Results: User experience

- **Questionnaire analysis:**
  - Wilcoxon matched-pairs signed ranks test.
  - Significant difference in confidence sub-scale ($p < 0.05$).
  - Higher confidence without feedback.
  - No significant difference in other 6 subscales and overall.
- **Feedback Preference:**
  - 97% preferred receiving feedback.



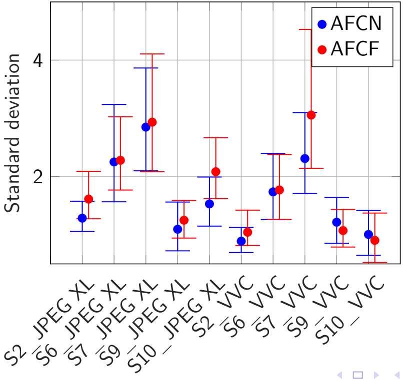
Results: Assessment of JND

- **H1**: Feedback does not influence the JND estimation.
- **Findings:**
  - No significant difference in JND estimation between feedback (AFCF) and no feedback (AFCN) conditions.
- **Conclusion:** H1 is not rejected.



Results:Precision of JND assessment

- **H2:** Feedback increases the precision of JND assessment.
- **Findings:**
  - Standard deviation of Weibull distributions used to measure precision.
  - Overlapping confidence intervals indicate no significant difference in precision between conditions.
- **Conclusion:** H2 is rejected.



Conclusion

Feedback impact:

- No bias in JND estimation.
- No significant effect on accuracy.
- No significant learning effect observed.
- Reduced response times.
- Almost all crowdworkers preferred receiving feedback.

Future works:

- Incorporating more gamification elements:
  - Leaderboards with ranking and scores
  - Progress bars
  - Competitive features

Access the dataset and subjective data here: