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

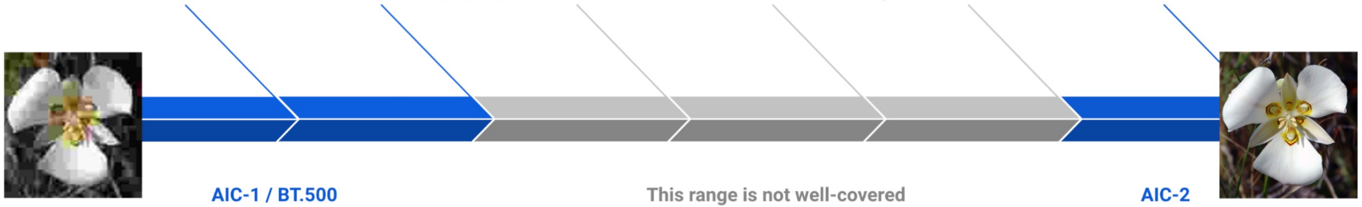
+

Shaolin Su, Oliver Wiedemann, Hui Men

VQEG June 2023

École
polytechnique
fédérale
de Lausanne

Subjective Visual Quality Assessment



AIC-1 / BT.500

Side-by-side comparison,
appeal-oriented impairment scale

This range is not well-covered

With BT.500, these are already "visually transparent", so indistinguishable.

With AIC-2, these are "clearly not visually transparent", so indistinguishable.

AIC-2

Flicker test,
fidelity-oriented,
visually lossless

ISO/IEC JTC 1/SC 29/WG 1
(ITU-T SG16)

Coding of Still Pictures

JBIG
Joint Bi-level Image
Experts Group

JPEG
Joint Photographic
Experts Group

The work of the JPEG AIC project produced a technical report, *Guidelines for image coding system evaluation* in ISO/IEC TR 29170-1:2017 and a standard, the Evaluation procedure for nearly lossless coding, in ISO/IEC 29170-2:2015.

JPEG AIC-3 Dataset

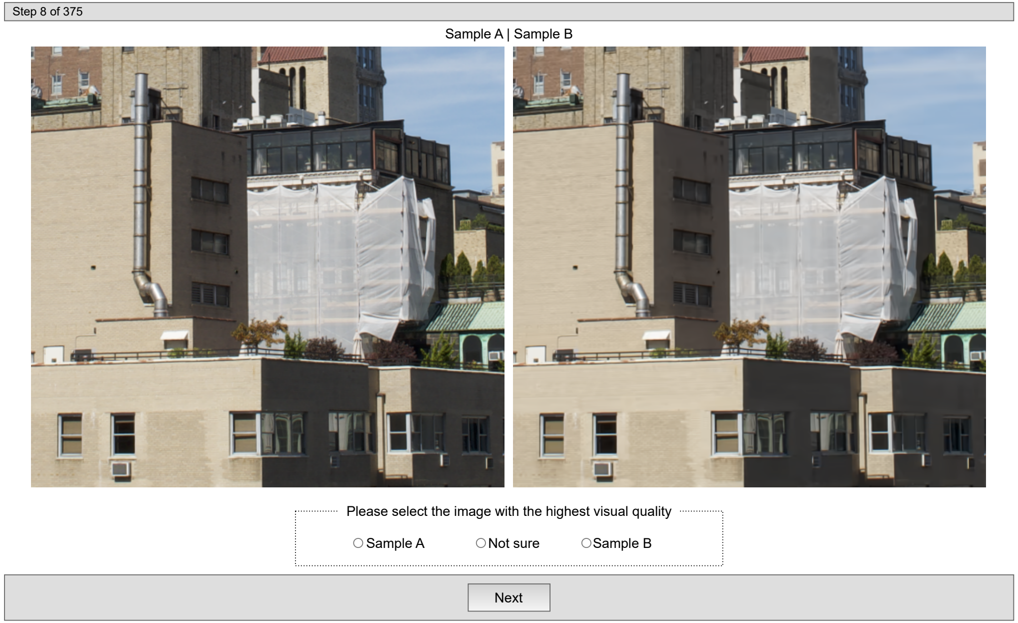
- **10** reference images, different **resolutions** and **content**
- Compression artifacts generated with **JPEG**, **JPEG 2000**, **HEVC Intra**, **VVC Intra**, and **JPEG XL** at multiple quality levels
- Visual quality range from **high to nearly visually lossless**
  - Selected through a **subjective**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



Step 8 of 375

Sample A | Sample B

Please select the image with the highest visual quality

○ Sample A ○ Not sure ○ Sample B

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

[1] H. Men, H. Lin, M. Jenadeleh, and D. Saupe, "Subjective image quality assessment with boosted triplet comparisons," IEEE Access, vol. 9, pp. 138 939–138 975, 2021.

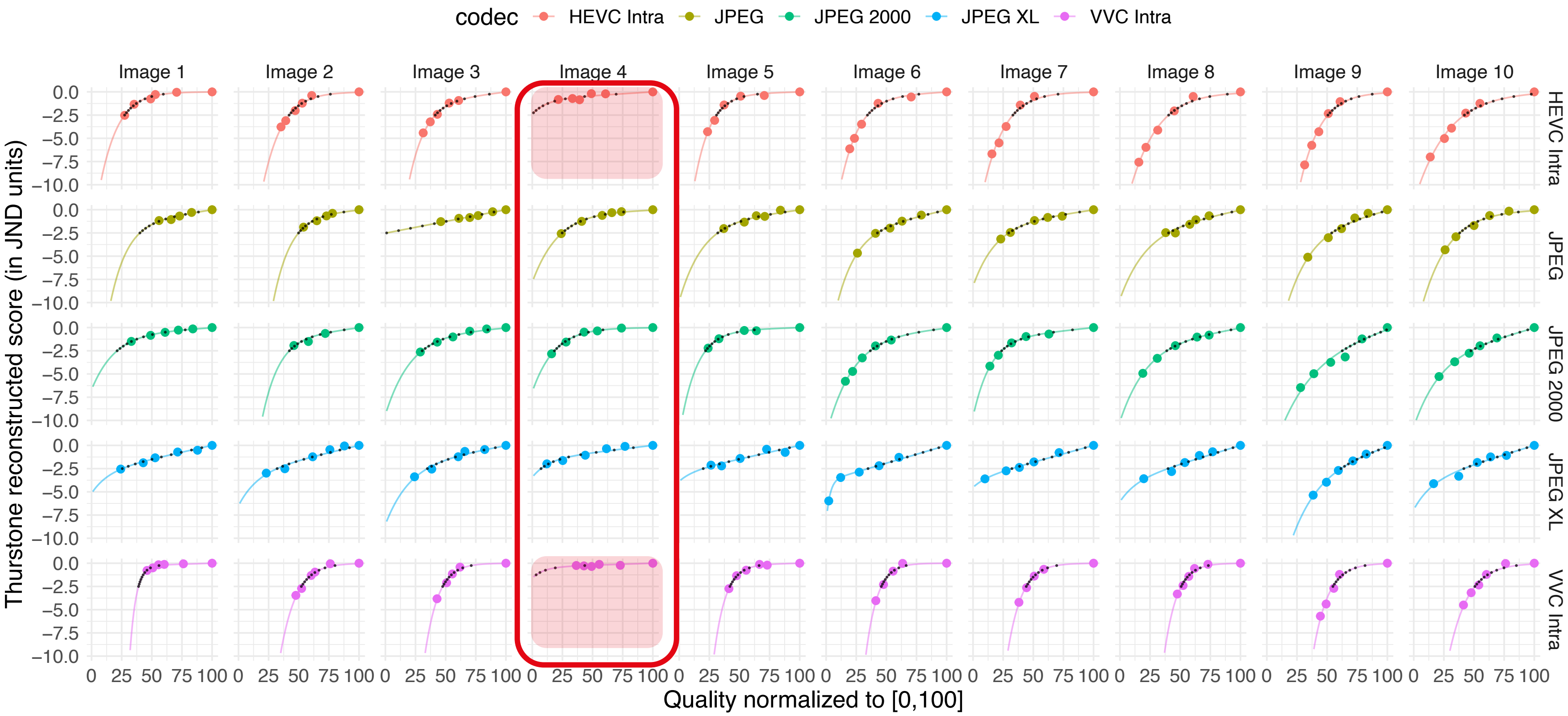
Statistical analysis

- 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

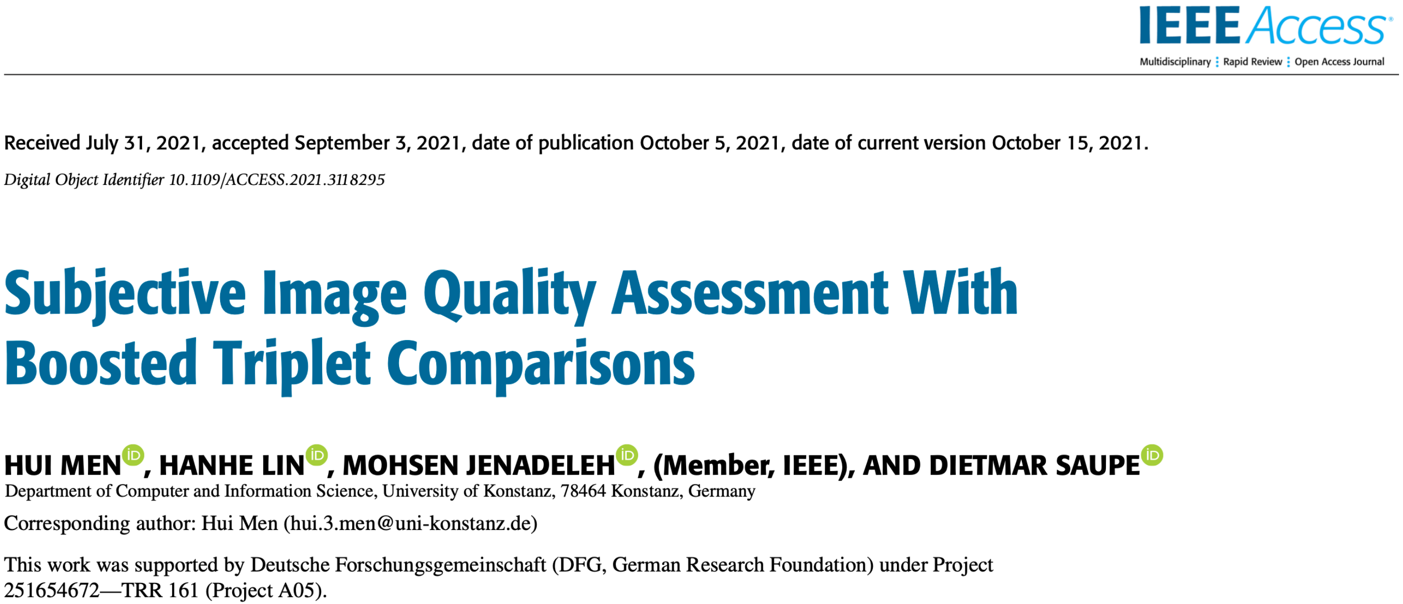


Fine-grained assessment of subjective quality of compressed images

Vlad Hosu[1], Mohsen Jenadeleh[1], Shaolin Su[1], Oliver Wiedemann, Hui Men,
Dietmar Saupe, University of Konstanz

[1]The first three authors contributed equally.

Our proposal: Boosted triplet comparison



IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

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

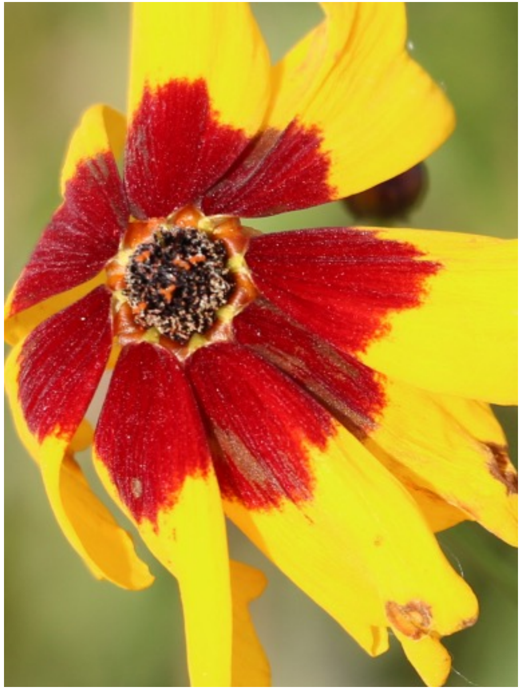
Subjective Image Quality Assessment With Boosted Triplet Comparisons

HUI MEN, HANHE LIN, MOHSEN JENADELEH, (Member, IEEE), AND DIETMAR SAUPE
Department of Computer and Information Science, University of Konstanz, 78464 Konstanz, Germany

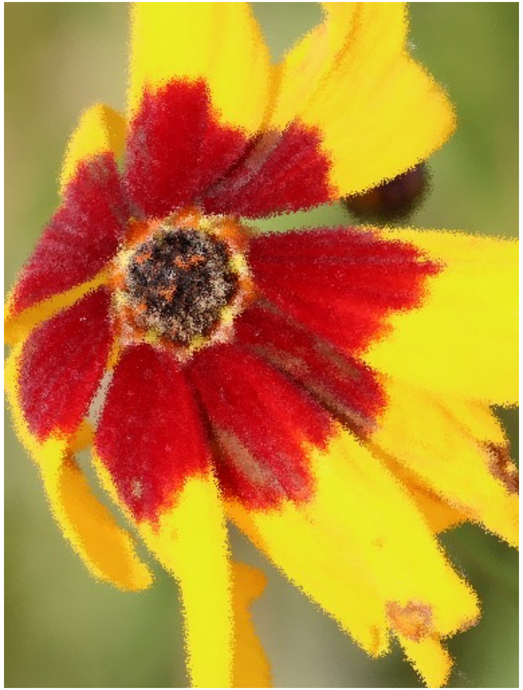
Corresponding author: Hui Men (hui.3.men@uni-konstanz.de)

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

Reference and Distorted Image



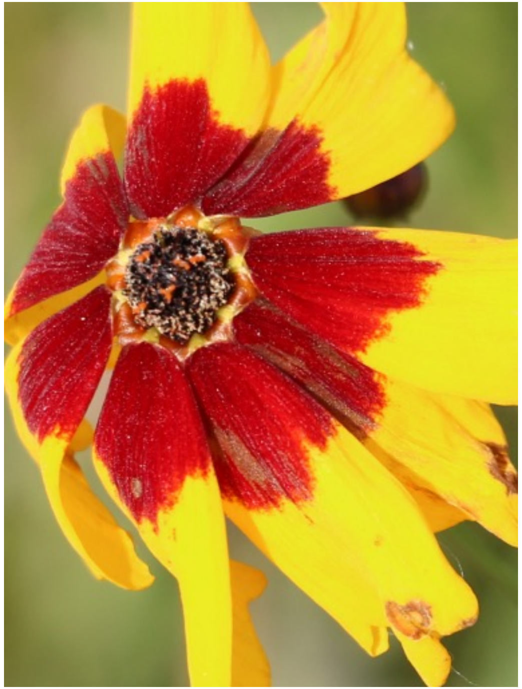
Ref.



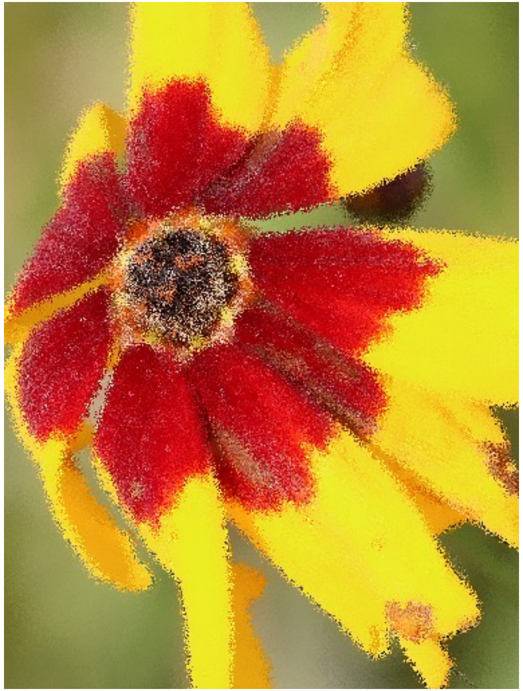
Orig. Dist.

Boosting (A)

$v' = v_{ref} + \alpha(v_{dist} - v_{ref})$ $(\alpha>1)$

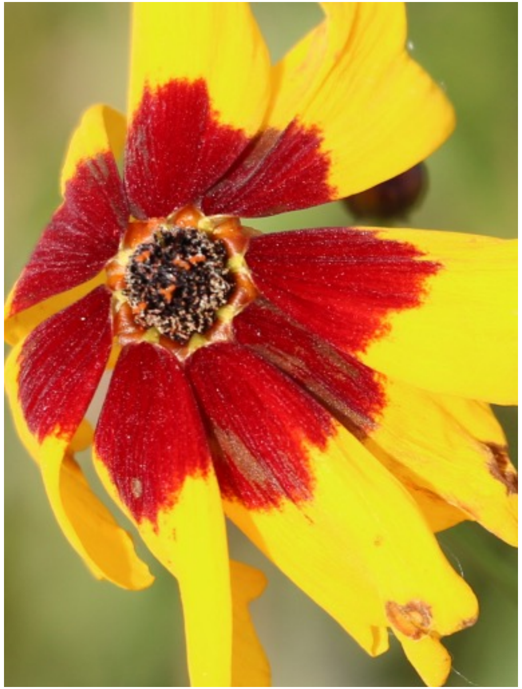


Ref.

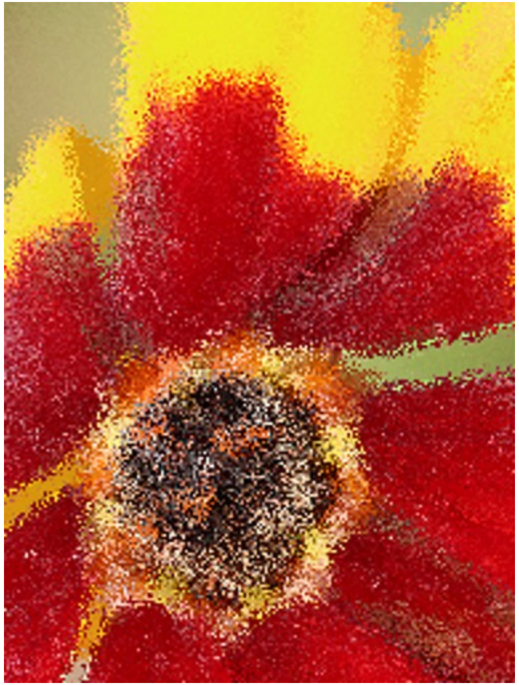


Amplification (A)

Boosting (A+Z)



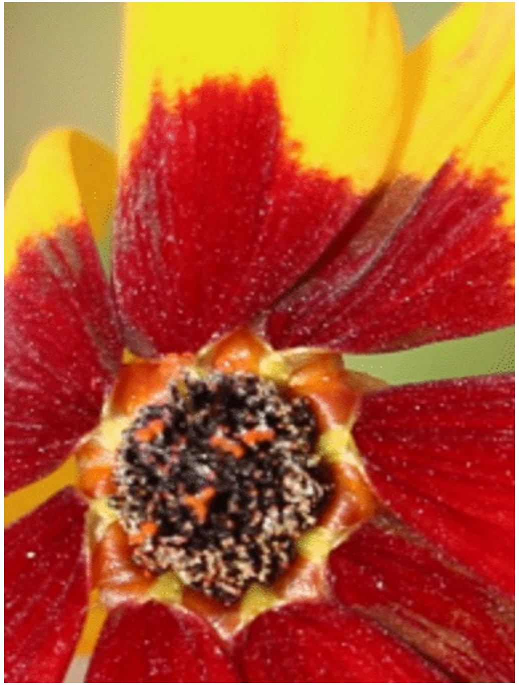
Ref.



Added Zoom (Z)

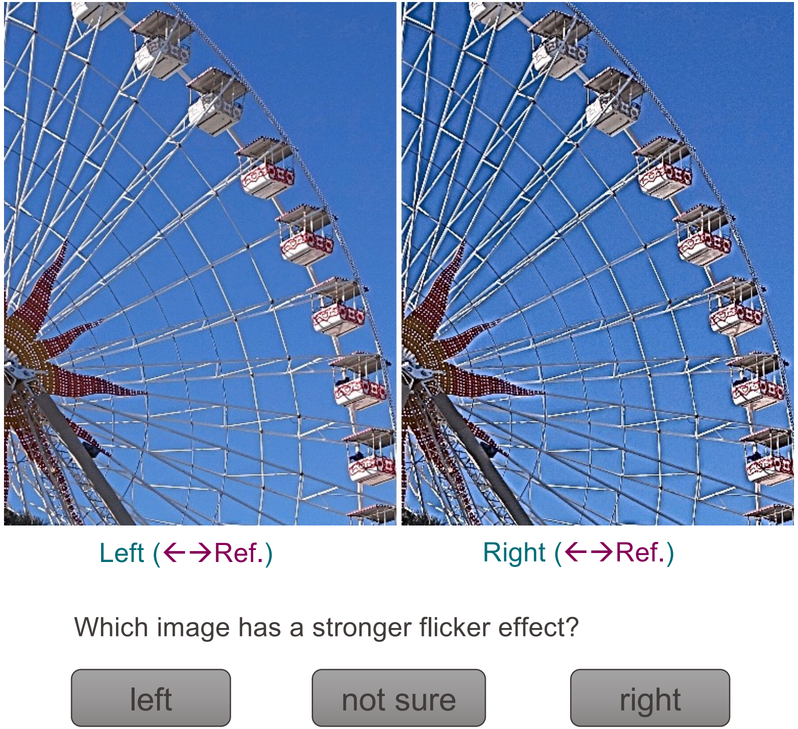
Boosting (A+Z+F)

Ref.



Added Flicker (F)

Comparison of two compressed images flickering w.r.t. source image

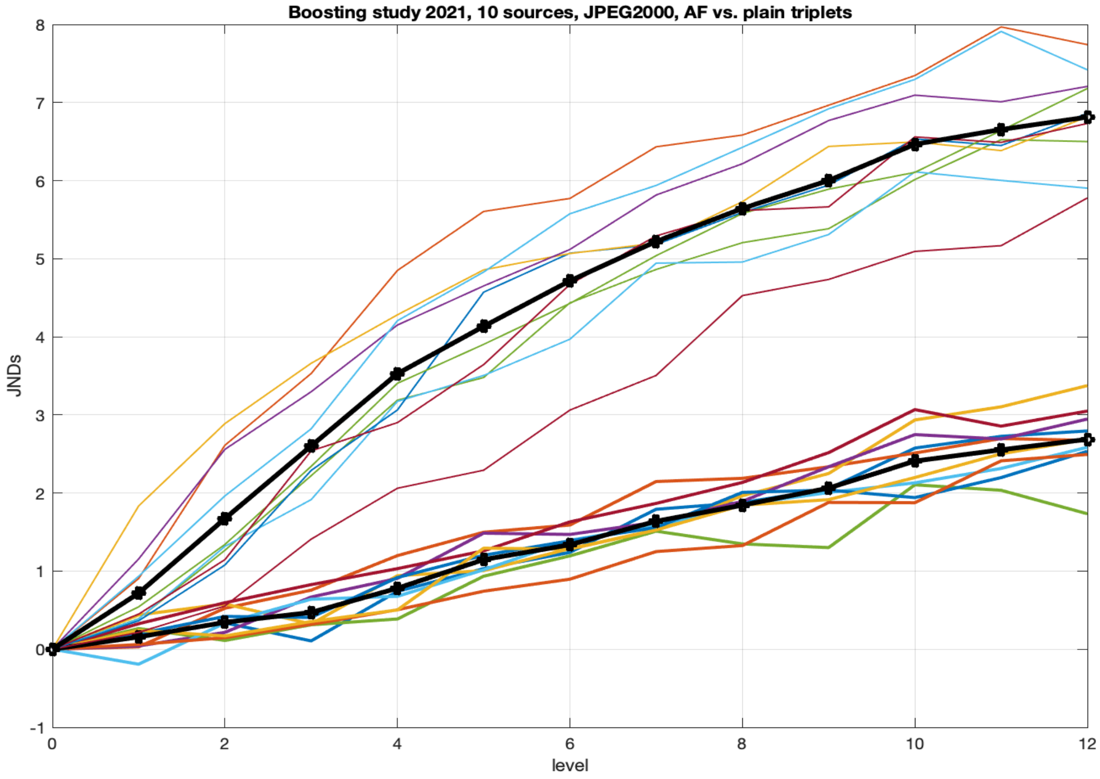


Left (←→Ref.)

Right (←→Ref.)

Which image has a stronger flicker effect?

- Impairment Scales for JPEG2000, A+F, 10 images and average



Application for JPEG AIC-3 dataset

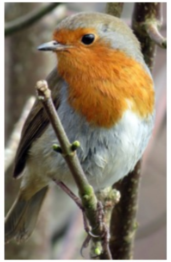
0

1

2

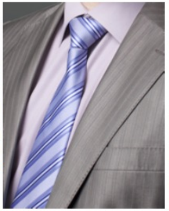
3

4



5

6



7

8

9

- 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

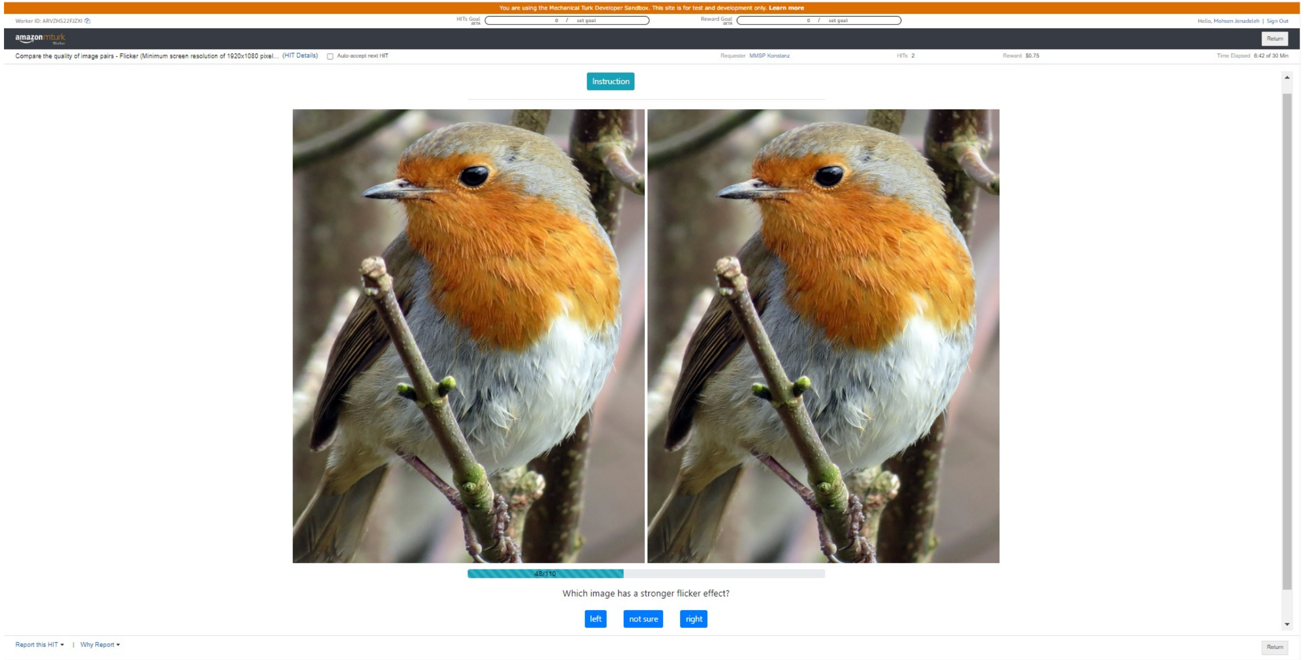
- Baseline triplets are (i,0,k)
  - Two images at levels i and k are compared with the source (level 0)
- Same-codec and cross-codec comparisons
- Selection of triplet comparisons:
  - Per sequence of 11 images: All 110 triplets (i,0,k) with i < k or k < i.
  - This makes 60*110 = 6600 same-codec triplets
  - 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

Zerman, E., Valenzise, G., & Smolic, A. (2019, June). Analysing the impact of cross-content pairs on pairwise comparison scaling. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*

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 assigments
  - 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



You are using the Mechanical Turk Developer Sandbox. This site is for test and development only. **Learn more**

Worker ID: ARVZH522FJZXI

HITs Goal BETA 0 / set goal

Reward Goal BETA 0 / set goal

Hello, Mohsen Jenadeleh | Sign Out

Compare the quality of image pairs - Flicker (Minimum screen resolution of 1920x1080 pixel... (HIT Details) □ Auto-accept next HIT

Requester MMSP Konstanz

HITs 2

Reward $0.75

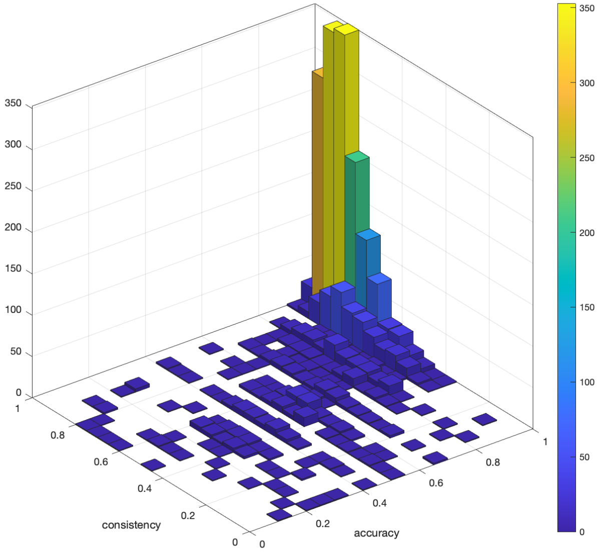
Time Elapsed 6:42 of 30 Min

48/110

Which image has a stronger flicker effect?

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)



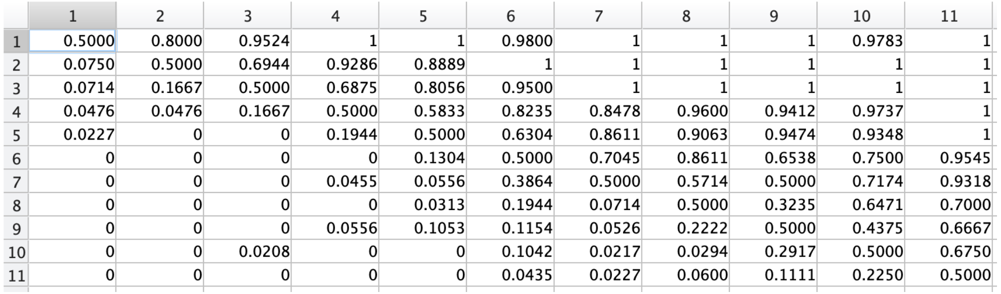
Accuracy and consistency for 2265 assignments

Data filtering and outlier removal

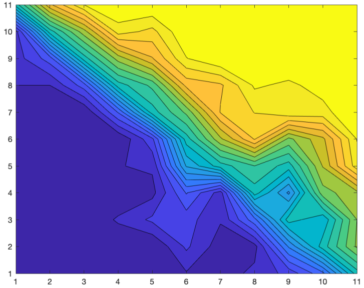
- Filtering and outlier removal at assignment level (110 triplets each)
- Assignments will be included if all of following hold:
  - Number of skipped questions <= 10
  - Accuracy >= 0.7
  - Consistency >= 0.6
- Iterative outlier removal for the remaining assignments based on negative log-likelihood (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

Example: Image 0, Codec 1

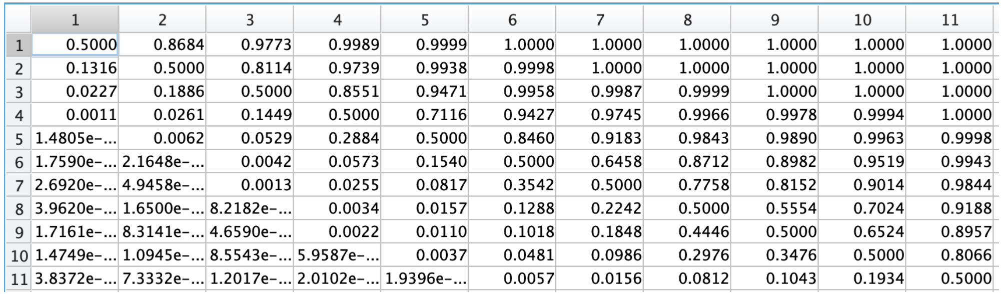
Empirical probabilities from experiment



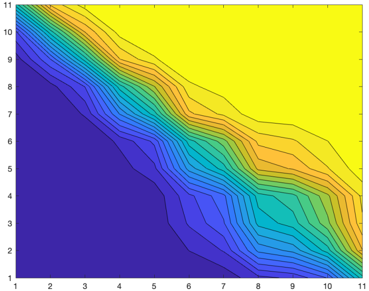
	1	2	3	4	5	6	7	8	9	10	11
1	0.5000	0.8000	0.9524	1	1	0.9800	1	1	1	0.9783	1
2	0.0750	0.5000	0.6944	0.9286	0.8889	1	1	1	1	1	1
3	0.0714	0.1667	0.5000	0.6875	0.8056	0.9500	1	1	1	1	1
4	0.0476	0.0476	0.1667	0.5000	0.5833	0.8235	0.8478	0.9600	0.9412	0.9737	1
5	0.0227	0	0	0.1944	0.5000	0.6304	0.8611	0.9063	0.9474	0.9348	1
6	0	0	0	0	0.1304	0.5000	0.7045	0.8611	0.6538	0.7500	0.9545
7	0	0	0	0.0455	0.0556	0.3864	0.5000	0.5714	0.5000	0.7174	0.9318
8	0	0	0	0	0.0313	0.1944	0.0714	0.5000	0.3235	0.6471	0.7000
9	0	0	0	0.0556	0.1053	0.1154	0.0526	0.2222	0.5000	0.4375	0.6667
10	0	0	0.0208	0	0	0.1042	0.0217	0.0294	0.2917	0.5000	0.6750
11	0	0	0	0	0	0.0435	0.0227	0.0600	0.1111	0.2250	0.5000



Model probabilities after MLE for Thurstonian model

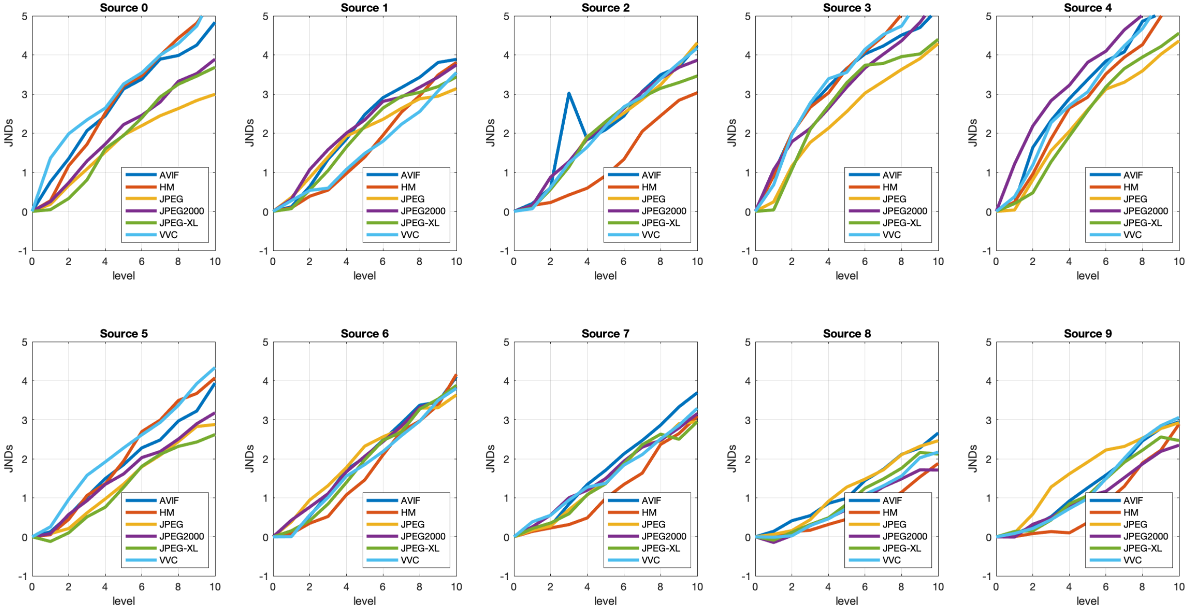


	1	2	3	4	5	6	7	8	9	10	11
1	0.5000	0.8684	0.9773	0.9989	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0.1316	0.5000	0.8114	0.9739	0.9938	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000
3	0.0227	0.1886	0.5000	0.8551	0.9471	0.9958	0.9987	0.9999	1.0000	1.0000	1.0000
4	0.0011	0.0261	0.1449	0.5000	0.7116	0.9427	0.9745	0.9966	0.9978	0.9994	1.0000
5	1.4805e-...	0.0062	0.0529	0.2884	0.5000	0.8460	0.9183	0.9843	0.9890	0.9963	0.9998
6	1.7590e-...	2.1648e-...	0.0042	0.0573	0.1540	0.5000	0.6458	0.8712	0.8982	0.9519	0.9943
7	2.6920e-...	4.9458e-...	0.0013	0.0255	0.0817	0.3542	0.5000	0.7758	0.8152	0.9014	0.9844
8	3.9620e-...	1.6500e-...	8.2182e-...	0.0034	0.0157	0.1288	0.2242	0.5000	0.5554	0.7024	0.9188
9	1.7161e-...	8.3141e-...	4.6590e-...	0.0022	0.0110	0.1018	0.1848	0.4446	0.5000	0.6524	0.8957
10	1.4749e-...	1.0945e-...	8.5543e-...	5.9587e-...	0.0037	0.0481	0.0986	0.2976	0.3476	0.5000	0.8066
11	3.8372e-...	7.3332e-...	1.2017e-...	2.0102e-...	1.9396e-...	0.0057	0.0156	0.0812	0.1043	0.1934	0.5000

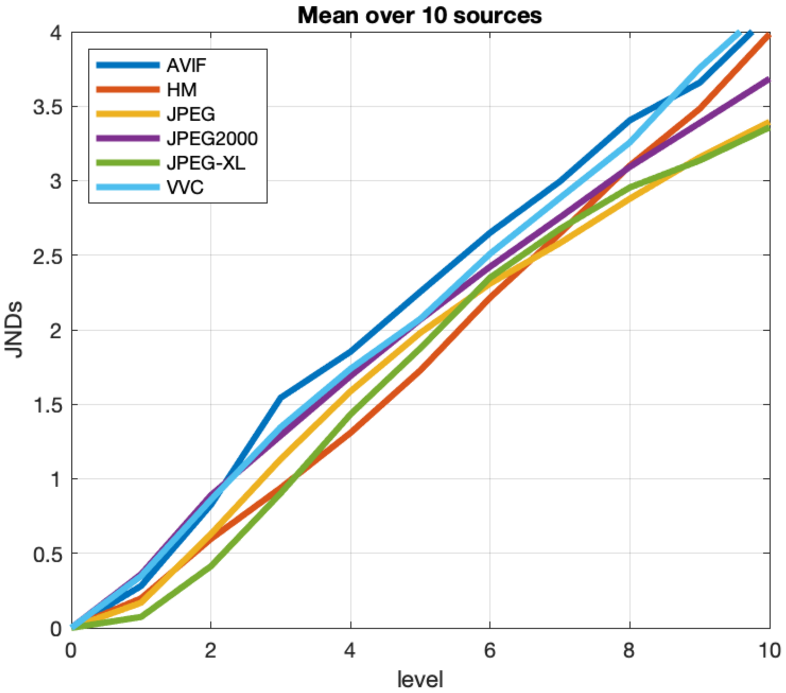


The empirical probabilities on the diagonal are not from the experiment. Stimuli were not compared with themselves. These values 0.5 are included only to help Matlab to create the heatmap correctly.

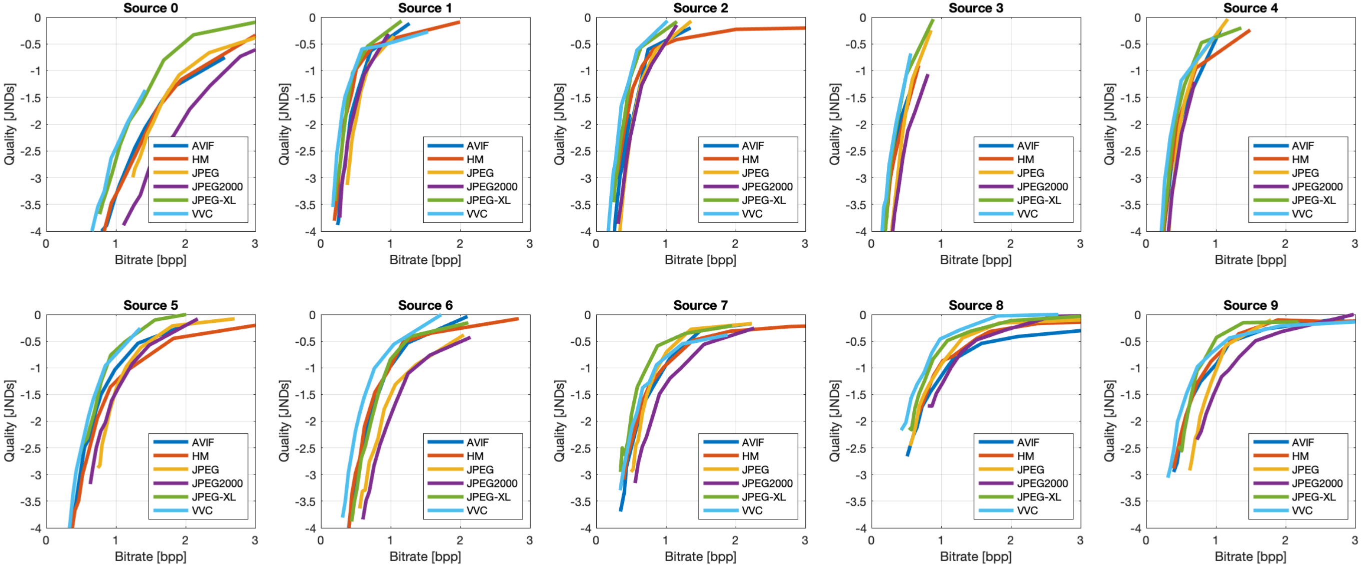
Perceived distortion vs distortion level



Perceived distortion vs. level: Summary



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