Psychometric scaling of TID2013 dataset

Aliaksei Mikhailiuk, María Pérez-Ortiz and Rafal Mantiuk
Department of Computer Science and Technology
University of Cambridge, Cambridge, UK
Email: am2442@cam.ac.uk

***Abstract*—TID2013 is a subjective image quality assessment dataset with a wide range of distortion types and over 3000 images. The dataset has proven to be a challenging test for objective quality metrics. The dataset mean opinion scores were obtained by collecting pairwise comparison judgments using the Swiss tournament system, and averaging votes of observers. However, this approach differs from the usual analysis of multiple pairwise comparisons, which involves psychometric scaling of the comparison data using either Thurstone or Bradley-Terry models. In this paper we investigate how quality scores change when they are computed using such psychometric scaling instead of averaging vote counts. In order to properly scale TID2013 quality scores, we conduct four additional experiments of two different types, which we found necessary to produce a common quality scale: comparisons with reference images, and cross-content comparisons. We demonstrate on a fifth validation experiment that the two additional types of comparisons are necessary and in conjunction with psychometric scaling improve the consistency of quality scores, especially across images depicting different contents.**

I. Introduction

One of the purposes of subjective image quality assessment is the construction of a metric for evaluating image degradation or enhancement as perceived by humans. Human judgments can be elicited via different types of experiments, such as explicit rating or comparative judgment approaches. Comparative judgment experiments have gained acceptance in subjective quality evaluation because of their simplicity. In a typical pairwise comparison experiment, observers are shown two images and are asked to select the one which appears to have a better quality. This approach greatly simplifies the task, as it avoids the need to rate each image explicitly. The quality scores can then be inferred from the matrix of pairwise comparisons by vote counting or psychometric scaling, with the latter having the advantage of providing a scale in which distances can be interpreted in terms of probability of better perceived quality.

One of the most recent and extensively evaluated image quality datasets (TID2013 [1], [2]) uses a crowd-sourcing experiment with pairwise comparisons in order to measure image quality. However, the quality scale was produced by dividing the number of votes given to each condition by the number of observers. The present paper shows the limitations of such an approach in comparison to the probabilistic framework that psychometric scaling provides. We construct the psychometric scale using the well-known Thurstone Case V model [3], and compare both strategies in simulated and real-world experiments. In addition, we complement TID2013 dataset with comparisons across different contents and with reference images, as neither were originally included in the dataset. Both are crucial for constructing a common quality scale. We show how inclusion of these additional comparisons changes the quality scale and how psychometric scaling helps to reduce the amount of inconsistencies. The new scale for TID2013 is available at the repository associated to this paper[1].

II. Related work

There are different methods for generating a scale of human preferences [4]. One of the ways to generate such a scale is through explicit rating, in which observers are asked to assign a numeric quality score to each object. All judgments are then averaged to produce mean opinion scores. Another way to generate the scale is to use comparative judgement experiments. In comparative judgement, observers are asked to compare two or more objects in terms of some quality criteria. Comparisons can be made in a pairwise or set-wise fashion, although pairwise comparisons are usually preferred for simplicity. All collected comparisons are later used to produce the final quality scores. Two commonly used methods to infer quality scores from pairwise comparisons are psychometric scaling and vote counts. Fig. 1 shows an example of both explicit ratings and pairwise comparisons with TID data.

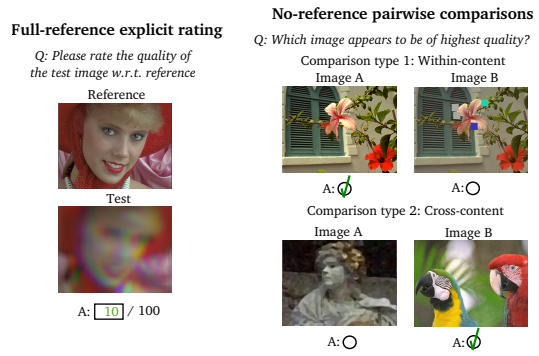

Fig. 1: Examples of full-reference explicit rating and no-reference pairwise comparison measurements.

A. TID2013

TID2013 presents 3000 distorted conditions (25 contents, 24 distortion types and 5 levels of distortions). Approximately 30

[1]https://doi.org/10.17863/CAM.21517

observers were involved in the measurement of every content, totalling more than 900 observers.

Each observer in one experiment only performed comparisons within one content, i.e. no cross-content comparisons were performed. In the pairwise comparison experiments the less distorted condition had to be chosen with the help of the reference displayed alongside. The pairs of conditions to compare were chosen following the Swiss chess system [5]. With this method, all conditions are compared the same predefined number of times. The first comparisons are chosen at random. In later stages, conditions are sorted based on the number of times they were previously selected by an observer, and conditions having similar quality compete in pairs.

The matrix of comparisons in TID2013 therefore has an incomplete unbalanced design with 25 disconnected components and no comparison to reference. Absence of cross-content comparisons and comparisons with the reference makes it impossible to construct a unified quality scale for all contents.

Subjective image quality scores presented in TID2013 are given in vote counts. Vote counts were obtained for each content separately by taking the total number of times a condition was selected as better and dividing by the number of observers. Every observer compared each condition within one specific content in nine pairwise comparisons, producing a scale between 0 and 9.

B. Comparative judgment and psychometric scaling

Comparative judgment approaches, such as pairwise comparisons, present numerous advantages over other types of explicit rating: i) they lead to a very simple experimental task and are therefore well suited for non-expert participants and crowd-sourcing experiments, ii) they avoid calibration issues frequently encountered with explicit rating [6], iii) they generally provide higher sensitivity and a lower measurement error [7], [8] and iv) the number of comparisons can be reduced using adaptive procedures [9], allowing for more evaluations for a given amount of time, effort and cost.

The results of a pairwise comparison experiment are usually represented using a count matrix $\mathbf{C}$, where each element c_{ij} measures the number of cases in which condition A_i has been selected over condition A_j.

1) Psychometric scaling: Suppose we aim to compare n conditions $A_1, \ldots, A_n$ with unknown underlying true quality scores $q = (q_1, \ldots, q_n)$, $q_i \in \mathbb{R}$. The role of psychometric scaling is to use aggregated comparison data to estimate scores $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_n)$ that approximate true quality q.

Estimated qualities $\hat{q}$ can be derived by recovering the underling relationships or distances between them: $d_{ij} = \hat{q}_i - \hat{q}_j$. Once these distances are inferred, the problem is transformed into one of dimensionality reduction.

Models for comparative judgment in psychometric scaling construct a quality scale in which distances d_{ij} between conditions A_i and A_j in the quality scale can be interpreted as a function of the probability of better perceived quality p_{ij}:

$$P(A_i \succ A_j) = p_{ij} = F(d_{ij}, s_{ij}), \tag{1}$$

where F is a cumulative distribution function of a random variable pertaining to the linear model chosen (e.g. Thurstone models assumes that F follows a Normal distribution with a standard deviation $\sigma_{ij} = s_{ij}$ [3], and Bradley-Terry assumes a logistic function with a parameter s_{ij} [10]). Models for psychometric scaling often use the Thurstone case V assumption and set σ_{ij} constant across all conditions. In practice, both Thurstone and Bradley-Terry present very similar solutions. In this paper we focus on the Thurstone case V model.

Probabilities p_{ij} can be estimated using the empirical information from the matrix $\mathbf{C}$:

$$\hat{p}_{ij} = \frac{c_{ij}}{c_{ij} + c_{ji}}, \quad i \neq j. \tag{2}$$

These empirical probabilities can be used to estimate distances between scores: $\hat{d}_{ij} = F^{-1}(p_{ij}, \sigma_{ij})$.

Psychometric scaling aims to find estimated scores $\hat{q}$ such that distances between scores closely resemble distances $\hat{d}_{ij}$. The simplest way to do so is to solve a least square optimization problem of the form [11]:

$$\arg\min_{\hat{q}_2,\ldots,\hat{q}_n} \sum_{i=1}^{n-1} \sum_{j=k}^{n} ((\hat{q}_i - \hat{q}_j) - \hat{d}_{ij})^2, \tag{3}$$

where q_1 is often set up as an anchor to 0, since scores are relative. This solution is simple but presents several problems: i) unanimous answers, in which all observers agree ($p_{ij} = 0$ or $p_{ij} = 1$), resulting in an infinite distance between A_i and A_j, and ii) confidence in the measurements is not considered.

A more elegant solution is provided by Maximum Likelihood using the Binomial distribution:

$$\arg\max_{\hat{q}_2,\ldots,\hat{q}_n} \prod_{i,j} \binom{n_{ij}}{c_{ij}} F\left(d_{ij}, \sigma_{ij}\right)^{c_{ij}} \left(1 - F\left(d_{ij}, \sigma_{ij}\right)\right)^{n_{ij} - c_{ij}}, \tag{4}$$

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

The choice of σ_{ij} determines the relationship between distances in the quality scale and probabilities of better perceived quality. Often, σ_{ij} is set to 1.4826, so that a distance of 1 unit can be interpreted as 75% of observers seeing a difference. These distance units are referred to as Just-Objectionable-Differences (JOD) [7] and are used throughout this paper.

2) Vote counts: If F in Eq. 1 is set to the uniform distribution, Eq. 3 corresponds to ranking conditions according to $\hat{q}_i = (1/n) \sum_{j=1}^{n} \hat{p}_{ij}$ or $\hat{q}_i = (1/n) \sum_{j=1}^{n} c_{ij}$. This last idea is usually referred to as vote counts — the number of times one condition was selected as better than any other condition. However this approach has limitations, i.e. $q_i \geq q_j$ doest not imply $p_{ij} \geq \frac{1}{2}$, when there are transitivity violations in the data or when conditions are not compared the same number of times.

In those situations, psychometric scaling algorithms are usually preferred. These algorithms will not only produce the correct ranking, but also capture the magnitude of the differences between conditions in a principled way. Moreover, vote counts do not explicitly account for the relative quality difference between the conditions, whereas psychometric

scaling infers the scores by considering relationships among all compared conditions.

In this regard, Zerman et al. compared the results of psychometric scaling and vote counts to the scores obtained in a direct rating experiment [12]. They showed that psychometric scaling scores are stronger related to rating scores than vote counts, confirming that quality magnitudes are better captured when pairwise comparison data is scaled.

III. Scaling simulation

In order to compare the scores produced by vote counts with those produced by psychometric scaling, we use a Monte Carlo simulation. Since ground truth is not available in TID2013, simulation of these experiments is necessary to draw conclusions about the consistency of both vote counts and psychometric scaling. Two types of data were used as the ground truth in our simulation: i) randomly generated quality scores within a fixed range and ii) TID2013 vote counts for content 1.

A. Simulation procedure

Simulation of an experiment was designed to mimic the Swiss chess system used in TID2013. In the simulation of a comparison between two conditions, every simulated observer chooses condition A_i over A_j with a probability defined by $P(A_i \succ A_j) \sim N(q_{A_i} - q_{A_j}, \sigma)$, where we set $\sigma = 1.4826$. Comparison matrices produced by every simulated observer are aggregated together. Vote counts are produced by summing elements along the rows of the resultant matrix i.e. the number of times every image was preferred divided by the number of observers. Psychometric scaling is produced using Maximum Likelihood estimation with the Thurstone Case V model and the Matlab code provided in [7]. The Spearman Rank Ordering Correlation Coefficient (SROCC) and Root Mean Squared Error (RMSE) are calculated for vote counts and psychometric scaling results. Simulation is repeated 1000 times and SROCC and RMSE values are averaged.

The simulation replicated the protocol from TID2013, i.e. every condition was compared 9 times, in 3 random and 6 sorted rounds using Swiss system, by 30 observers. True quality scores were set to vary between 0 and 9 in the experiment with randomised data (similar to the scale in TID2013) and true quality scores were assigned random vote counts from content 1 in the simulation with TID data.

B. Simulation results

The results of the simulation are depicted in Fig. 2. The positive difference between psychometric scaling and vote counts in SROCC and negative in RMS, regardless of the number of conditions, indicates that psychometric scaling consistently outperforms vote counts in estimating both the ranking and the scale. The difference in SROCC between psychometric scaling and vote counts increases with the number of conditions, and so psychometric scaling is preferable as the number of conditions increases.

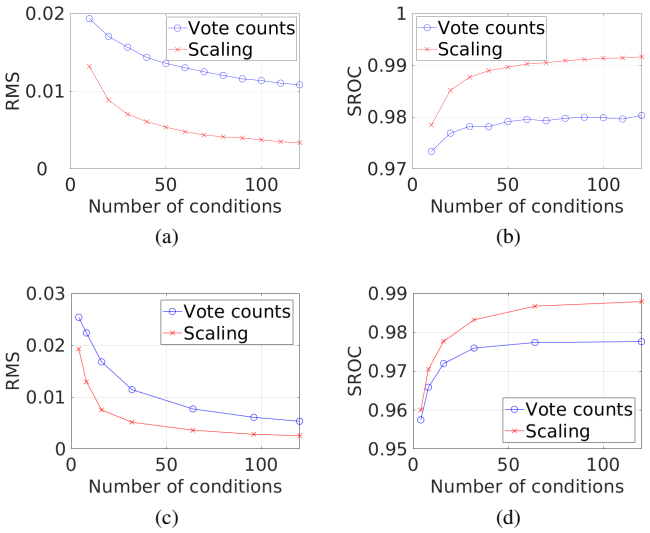

Fig. 2: (a), (b) Simulation of the experiment with TID data. (c), (d) simulation of the experiment with random data.

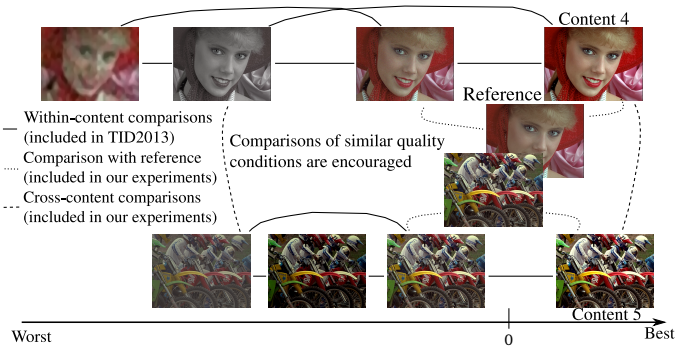

Fig. 3: Representation of different types of comparisons necessary to position all contents on a common quality scale.

IV. Psychometric scaling of TID2013

The experimental procedure in TID2013 presents two important limitations. Firstly, although reference images were used to help observers to choose between distorted images, they were never compared with distorted conditions. Existence of a common quality anchor for every content, i.e. a reference image, is necessary for constructing a fully connected graph of comparisons. Secondly, comparisons across different contents were not performed. Without cross-content comparisons, contents cannot be accurately scaled [12]. Therefore original TID2013 scores cannot be compared across different contents. These required types of comparisons are shown in Figure 3.

A. Experimental setup

We extend the data collected in TID2013 with five additional experiments. The first experiment was used to include reference images. The next three include cross-content comparisons and further comparisons to the reference to improve the scale. The last experiment is a validation experiment, used

to evaluate the quality of scales produced by vote counts and psychometric scaling. We later use comparisons from all five experiments to construct the final scale. Each experiment had 10 participants, with each participant completing 300 trials. Overall additional 15,000 comparisons were collected. For the design of the first four experiments, we take into account the fact that it is often more informative to compare conditions that are of similar quality (which is the main motivation for the use of the Swiss system in TID2013). We asked observers to select the better quality image among distorted ones. The order of comparisons in every experiment was randomized. We ensure that ITU recommendations are met and that the time for performing one experiment does not exceed 30 mins, so as to prevent observer tiredness from influencing the experiment outcomes.

1) Inclusion of reference: For the first experiment, we scale each content separately and compare each reference image to four conditions (within the same content) that have the best quality score, so as to gain the most information from the comparisons [7]. This produces $25 * 4 = 100$ comparisons, which allow the inclusion of reference images in the quality scale. Each measurement is repeated 3 times by each of the 10 observers. We use the newly collected data to scale the dataset, where we assume all reference images to have a common quality score of zero. In this scale conditions can have both positive and negative scores, where a positive score is attributed to image enhancement and negative to image distortions.

2) Inclusion of cross-content comparisons and scale refinement: For the next three experiments, we include 300 more comparisons, most of which were cross cross-content, i.e. 40 comparisons to the reference, 240 cross content comparisons and 20 within content comparisons. After each experiment we rescaled the data and used produced results to select comparisons for the next experiment.

3) Validation experiment: In order to compare the consistency of both scales (VC and JOD), we conduct an additional experiment. Using both scales (as represented in the left part of Fig. 4), we can find the cases in which VC and JOD differ the most. This strategy is referred to as Maximum differentiation (MAD) competition [13] and it is used to compare subjective and objective image quality metrics. We select 150 pairs of conditions A_i, A_j for which JOD is as different as possible and VC scores are as close as possible i.e. $argmax_{ij}(|\text{JOD}_i - \text{JOD}_j| - |\text{VC}_i - \text{VC}_j|)$. And similarly for A_i, A_j for which VC are different and JOD are similar. To promote diversity we allow each content to participate only in 50 comparisons. Overall we select 300 pairs of images with 150 images in each group and ask 10 observers to perform a pairwise comparison experiment.

B. Results and discussion

We firstly compute the correlation between the probability of an image being better (inferred from the validation experiment following Eq. 2) and the difference in quality scores in both VC and JOD scales for selected validation images. The SROCC between $\text{VC}_i - \text{VC}_j$ and p_{ij} is 0.52 and the SROCC between $\text{JOD}_i - \text{JOD}_j$ and p_{ij} is 0.69, indicating that the output of our new JOD scale reflects better image quality. After this validation experiment, we can include the data from the last experiment into our psychometric scale. Now, the SROCC is 0.84, meaning that psychometric scaling can successfully include the information of collected comparisons, thus the scale can be further improved in the presence of inconsistencies. The SROCC is, however, still far from 1, this is because the psychometric scaling finds the best one-dimensional scale taking into account all relationships in the data, which might not be optimal for a subset of selected conditions.

Fig. 5 can also be used to validate both scales. This figure shows the histograms of probabilities $P(A_i \succ A_j | \text{SC}_i \gg \text{SC}_j)$, which describe the percentage of observers selecting one condition over another given scores $\text{SC}_i \gg \text{SC}_j$ in either JOD or VC scales. The histogram on the left shows the first 150 pairs of conditions A_i and A_j for which $\text{JOD}_i \gg \text{JOD}_j$ and $\text{VC}_i \approx \text{VC}_j$. Similarly the histogram on the right shows the last 150 pairs of conditions A_i and A_j for which $\text{JOD}_i \approx \text{JOD}_j$ and $\text{VC}_i \gg \text{VC}_j$. Ideally, the best scale should present a large $P(A_i \succ A_j)$ for most cases. The concentration of counts around $P(A_i \succ A_j) \approx 1$ is higher for JOD, the counts for the VC scale gradually increase, whereas for the JOD the change is abrupt, with a sharp rise of the counts. The number of counts with high probability $P(A_i \succ A_j) > 0.7$ is also greater for JOD scaled data.

Another way of evaluating the consistency of both scales is to compute the log likelihood of observing the data collected in the validation experiment for two hypotheses: $\text{A}_i \approx \text{A}_j$, and hence $P(A_i \succ A_j) = 0.5$ and $\text{A}_i \gg \text{A}_j$, and hence $P(A_i \succ A_j) = 0.9$ (not 1 since the average distance between pair of images chosen is 2 JOD, corresponding to $\approx 90\%$ of observers choosing one image over another). For image pairs in which $\text{JOD}_i \gg \text{JOD}_j$ and $\text{VC}_i \approx \text{VC}_j$ the log likelihood of data under the assumption $\text{A}_i \approx \text{A}_j$ is -672.98 and under the assumption $\text{A}_i \gg \text{A}_j$ -393.34, suggesting this that the JOD scale better explains the validation data. Whereas for image pairs in which $\text{VC}_i \gg \text{VC}_j$ and $\text{JOD}_i \approx \text{JOD}_j$ the log likelihood of data under the assumption $\text{A}_i \approx \text{A}_j$ was -546.98 and under the assumption $\text{A}_i \gg \text{A}_j$ is -543.47. This indicates that conditions far apart in the VC scale are equally likely to come from both hypotheses.

Fig. 4 shows the relationship between vote counts and JOD scale produced by psychometric scaling. The plot in the left part of Fig. 4 shows the relationship after including comparisons from the first four experiments. It can be seen that there are some cases, e.g. content 5 and 8, which are consistently ranked better on the JOD quality scale than the rest, and others, such as contents 4, 10 and 12, which are consistently ranked worse. There are several reasons for this effect. First of all only a small number of cross-content experiments were performed, and the selection of compared conditions might not be sufficient to accurately capture all variations in the quality. Secondly, annoyance caused by different distortions

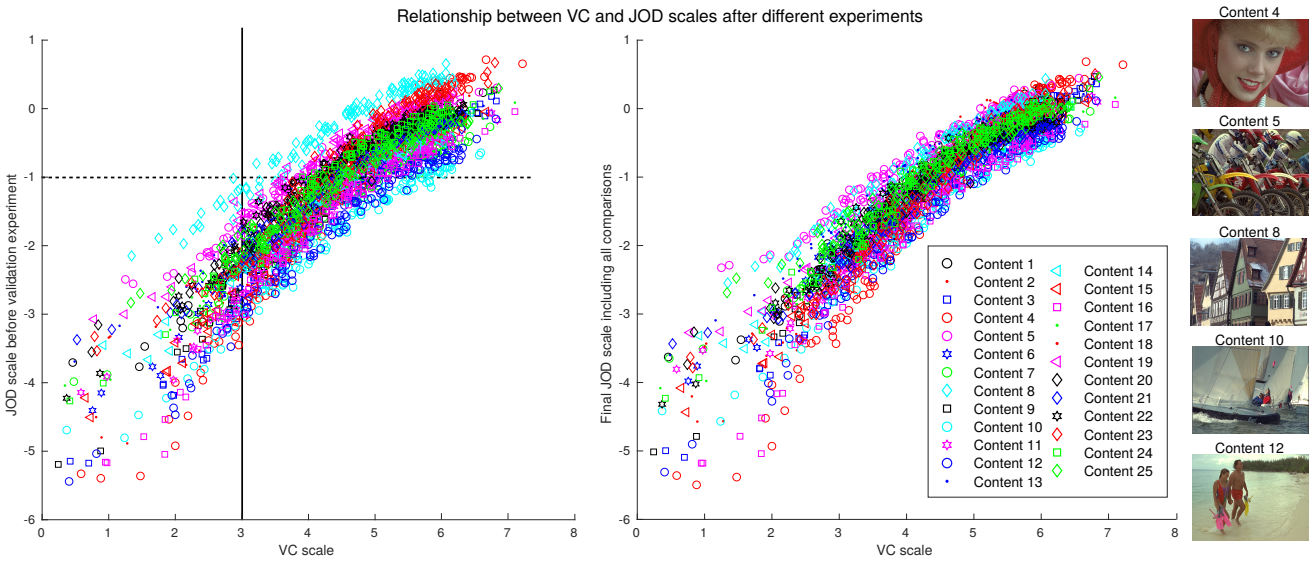

Fig. 4: Relationship between VC and JOD scales when adding new data. Left part shows JOD scale using the data from the first 4 experiments and a representation of the maximum differentiation competition used to select the validation comparisons. One scale is kept constant, while we chose images that in the other scale greatly differ in quality. The final scale after including all the comparisons from all five experiments can be seen in the right part.

is conditional on the content to which they are applied, for example these are more noticeable on human faces. The relationship between VC and the final psychometric scaling, which includes the data from all experiments, is shown in the right part of Fig. 4. Here the contents are more mixed in the scale. Both scales have a large positive correlation for conditions within the same content. We hypothesize that this is because the original TID dataset contains many more comparisons than the ones we collected, thus having a greater impact on psychometric scaling. The overall SROCC between VC and the final JOD scale is 0.9407.

A selection of pairs of images used for the validation experiment which have the largest inconsistencies in both scales is plotted in Fig. 6. Interestingly, most of the cases are cross-content and cross-distortion. The first two rows of pairs in Fig. 6 show obvious failures of the VC scale, which are solved by the JOD scale. The last two rows show failures in the JOD scale, which are however, less obvious.

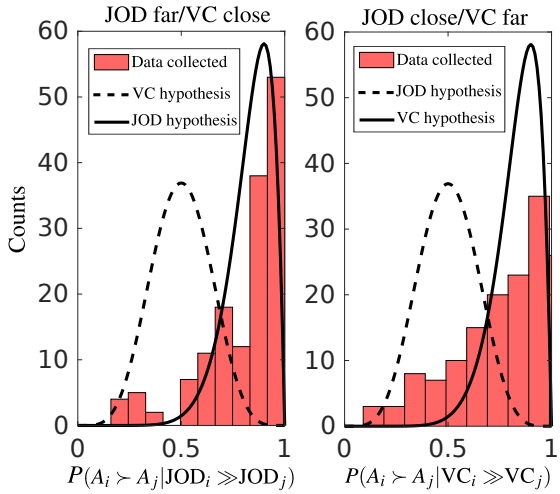

Fig. 5: Histograms of results of the validation experiment. The horizontal axis is the probability of image A_i being better than A_j, p_{ij}, inferred from the validation experiment.

C. Limitations

TID2013 comprises more than 400,000 comparisons, however, the dataset can still be improved in several ways: (i) Unanimous answers (in which all observers agree) represent 56% of comparisons in the dataset and these may introduce a bias in the scaling, as no upper bound is imposed on the distance between compared conditions [7]. The majority of these answers are due to some conditions being compared only once by one observer because of the use of the Swiss system and (ii) A greater number of additional comparisons would improve the scaling further.

The psychometric scaling used in this paper has a number of limitations: (i) observers or repetitions effects are not considered (specific scaling models can can account for this); (ii) the model represents quality in a one dimensional scale, however due to transitivity violations present in the data, one-dimensional scales might not be enoguh to represent quality scores [14] and (iii) case V of Thurstone model was used, however quality scores may have different variances.

V. Conclusions

In this paper we investigated how quality scores change when they are computed using psychometric scaling instead of averaging vote counts and extended the TID2013 dataset by conducting five additional experiments. We showed that

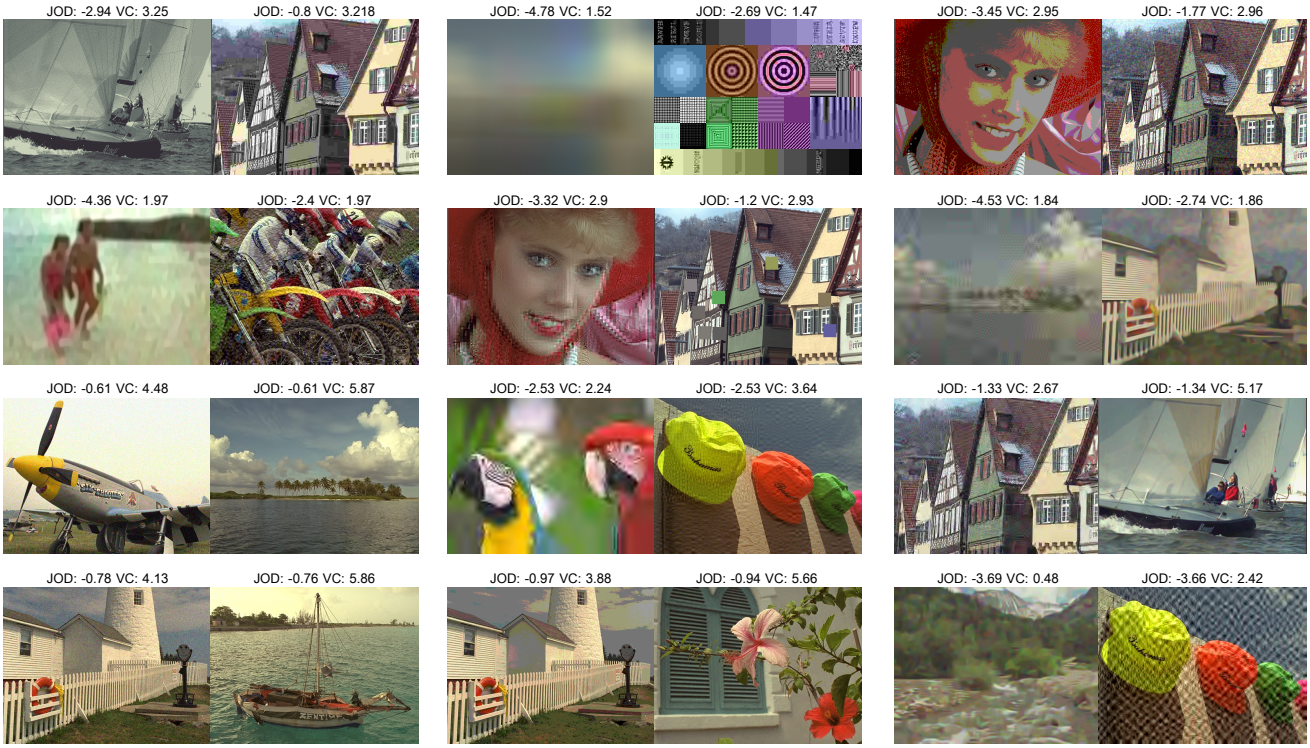

Fig. 6: Representation of comparisons where VC and JOD scales differ the most and empirical probabilities are unanimous (all observers agreed). In each pair the image on the left is the one which was not chosen by any of the observers and on the right is the one chosen by all observers. First six pairs represent cases where VC failed to correctly rank conditions (but JOD succeeded) and last six comparisons depict failure cases in JOD (where VC ranking succeeds).

psychometric scaling produces more accurate results than vote counting in a simulated experiment, especially as the number of conditions in the experiment increases. We also demonstrated that the additional set of comparisons and psychometric scaling improve the consistency of quality scores of the TID2013.

As future work, active sampling methods could be used to collect more data and reduce the number of comparisons while maximising the information gain. We also plan to use the newly scaled dataset to re-evaluate objective quality metrics.

Acknowledgment

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725253–EyeCode). This work was also supported by the EPSRC research grant EP/P007902/1.

References

[1] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[2] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, and Benoit, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57 – 77, 2015.

[3] L. L. Thurstone, "A law of comparative judgment." *Psychological Review*, vol. 34, no. 4, pp. 273–286, 1927.

[4] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: A survey," *CoRR*, vol. abs/1406.7799, 2014.

[5] L. Csató, "Ranking by pairwise comparisons for swiss-system tournaments," *Central European Journal of Operations Research*, vol. 21, no. 4, pp. 783–803, 2013.

[6] K. Tsukida and M. R. Gupta, "How to Analyze Paired Comparison Data," Department of Electrical Engineering University of Washington, Tech. Rep. UWEETR-2011-0004, 2011.

[7] M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," *CoRR*, 2017.

[8] N. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright, "Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, vol. 38. PMLR, 2015, pp. 856–865.

[9] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Computer Graphics Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[10] R. A. Bradley and M. E. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, vol. 39, no. 3/4, p. 324, dec 1952.

[11] P. G. Engeldrum, *Psychometric scaling: a toolkit for imaging systems development*. Imcotek Press, 2000.

[12] E. Zerman, V. Hulusic, G. Valenzise, R. K. Mantiuk, and F. Dufaux, "The relation between MOS and pairwise comparisons and the importance of cross-content comparisons," in *Proc. of Human Vision and Electronic Imaging*, 2018.

[13] W. Zhou and S. Eero, "Maximum differentiation competition: A methodology for comparing quantitative models of perceptual discriminability," *Journal of Vision*, vol. 5, no. 8, 2005.

[14] M. Cattelan, "Models for paired comparison data: A review with emphasis on dependent data," *Statistical Science*, vol. 27, no. 3, pp. 412–433, 2012.