An Image Quality Dataset with Triplet Comparisons for Multi-dimensional Scaling

Mohsen Jenadeleh*, Frederik L. Dennig*, Rene Cutura†, Quynh Quang Ngo†,
Daniel A. Keim*, Michael Sedlmair†, Dietmar Saupe*
*Department of Computer and Information Science, University of Konstanz, Konstanz, Germany
†VISUS, University of Stuttgart, Stuttgart, Germany
{mohsen.jenadeleh, frederik.dennig, daniel.keim, dietmar.saupe}@uni-konstanz.de,
{rene.cutura, quynh.ngo, michael.sedlmair}@visus.uni-stuttgart.de

***Abstract*—In the early days of perceptual image quality research more than 30 years ago, the multidimensionality of distortions in perceptual space was considered important. However, research focused on scalar quality as measured by mean opinion scores. With our work, we intend to revive interest in this relevant area by presenting a first pilot dataset of annotated triplet comparisons for image quality assessment. It contains one source stimulus together with distorted versions derived from 7 distortion types at 12 levels each. Our crowdsourced and curated dataset contains roughly 50,000 responses to 7,000 triplet comparisons. We show that the multidimensional embedding of the dataset poses a challenge for many established triplet embedding algorithms. Finally, we propose a new reconstruction algorithm, dubbed logistic triplet embedding (LTE) with Tikhonov regularization. It shows promising performance. This study helps researchers to create larger datasets and better embedding techniques for multidimensional image quality. The dataset includes images and ratings and can be accessed at https://github.com/jenadeleh/multidimensional-IQA-dataset/tree/main.**

***Index Terms*—multidimensional image quality assessment, triplet comparison, image quality dataset**

I. Introduction

Perceived image and video quality are usually expressed as a scalar, one-dimensional variable. If the subjective evaluation is based on the absolute category rating (ACR) or a visual analog scale (VAS), the mean opinion score (MOS) is given. For paired comparisons or comparisons between more than two stimuli, a reconstruction algorithm generates quality scores on a latent scale based on an underlying statistical model.

However, as early as 1990, at the beginning of research into the evaluation of image/video quality, an ITU report proposed investigating perceived quality in a multidimensional framework [1]. The aim was to examine the relationships between objective and perceptual parameters and to relate the perceptual dimensions to quality judgments and viewer satisfaction.

Some studies had been carried out even earlier, around 1980 [2], [3], but few have followed since then. In these studies, perceptual speech and visual quality was assessed by difference scaling [3]–[7]. Streijl et al. [8] gave a comprehensive overview of the field in 2014. They pointed out: "Although there has been [...] research into the multiple dimensions of quality, the use of multi-dimensional models for quality assessment for the various modalities is still relatively immature." We think that today, ten years later, this still is the case.

One of the reasons for this is that the ITU standards refer to multidimensional quality assessment only by a method called 'direct scaling' [9]–[11]. In this method, several descriptive quality scales such as audio discontinuity and noisiness, compression distortion or image blur are evaluated separately [12]–[16]. However, the original purpose of multidimensional scaling is to discover and quantify such dimensions in a perceptual space.

Several multidimensional reconstruction algorithms for datasets of triplet responses have recently been developed and analyzed [17]. However, none of these have been applied to image/video quality assessment. In our work, we use the following triplet embedding methods: GNMDS [17], STE & t-STE [18], and CKL [19].

We also propose a new embedding algorithm based on a maximum likelihood estimation (MLE) of a logistic probability model. It includes a regularization term that integrates prior knowledge: The reconstructions for a sequence of images with increasing degrees of distortion of the same distortion type should ideally form a sequence of points lying on a curve. Therefore, penalizing a large curvature can improve the multi-dimensional embeddings.

Lastly, there is a lack of appropriately designed datasets. Here, we created such a dataset based on triplet comparisons. It facilitates pilot experiments to challenge and test multi-dimensional reconstruction methods.

Our main contributions are: (1) Our annotated dataset of 84 distorted images with 49,693 responses to 6,947 triplet comparisons. (2) A critical comparison of the performance of existing triplet embedding methods applied to the dataset, showing their potential and challenges. (3) The logistic triplet embedding (LTE) with Tikhonov regularization that greatly improves the quality of multidimensional scaling for our dataset.

II. Dataset of triplet responses for multidimensional image quality assessmemt

A. Source and test images

For our stimuli, we used a subset of KonFiG-IQA dataset [20]. In this dataset, the authors provided 10 source images, each accompanied by distorted versions at 12 levels of distortion for each of the seven selected distortion types

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161, projects A03 and A08, and DFG – Project ID 496858717, titled "JND-based Perceptual Video Quality Analysis and Modeling".

(high sharpening, motion blur, lens blur, jpeg2000, reference, jitter, color diffusion, and multinoise). The source images, with a resolution of 512 × 384 pixels, were cropped from the images in the MCL-JCI dataset [21].

To keep the size of subjective study manageable, we chose SRC31 and all its compressed versions for this study, i.e., 85 images (1 source and $7 \times 12 = 84$ distorted images).

B. Triplet comparisons

From all possible 85^3 triplet questions, we have selected the most informative ones, namely those that compare a pivot image with test images that are likely to be perceptually close to the pivot. For this purpose, we let each of the 85 images be the pivot in 78 triplets. The other two images in these triplets were chosen as follows. From all 85 images, we took the 13 images with the largest (finite) peak signal-to-noise ratio (PSNR) w.r.t. the pivot image, and used all $13 \cdot 12/2 = 78$ pairs as test images for the pivot. The order in each pair was randomized. This yielded 6,630 triplets for our study questions.

C. Crowdsourcing study

To conduct the subjective experiment on the Amazon Mechanical Turk (MTurk) platform, we used triplets for study questions and generated additional ones for trap questions. We published 20 questions per each Human Intelligence Task (HIT), consisting of 18 study and 2 trap questions. Therefore, for 6,630 study and 738 trap questions, we had 368 HITs of 20 questions each and one HIT of 8 questions (6 study and 2 trap questions). We posted each HIT with 9 assignments for 9 unique crowdworkers. Altogether, this resulted in 66,312 questions. The order of the trap and study questions was randomized for each worker to reduce bias.

Requirements for workers to participate included having completed at least 100 HITs in previous work on MTurk, with a 95% approval rate, as well as using a PC or laptop with a screen resolution of at least 1366 × 768 pixels to properly display the web interface with images and using a Google Chrome browser.

Participants compared the images on the left and right, choosing the one they perceived as more similar to the pivot image in the middle. They could also select "not sure". Fig. 1a shows a triplet comparison example. Including the "not sure" option decreased mental load and improved data fit in models, as found in [22], and has been used in some visual quality assessment experiments [20], [23]–[25].

The workers had 5 seconds to inspect and decide on the images. If they did not respond within the 5-second display time, the image would be hidden, and a gray page would appear, giving the workers an additional 3 seconds to answer. If the crowdworker failed to answer, the response would be labeled as "undecided".

The Institutional Review Board of the University of Konstanz ethically approved the experimental procedures and protocols used in this study.

D. Data cleansing

A total of 264 workers participated in our subjective experiment. We published 369 HITs, each assigned nine times, resulting in 3,321 assignments. We excluded 818 assignments due to either the failure of workers to correctly answer a trap question or their inability to answer three or more questions within the 8-second response time. From the remaining approved assignments, 49,952 responses were obtained. Of these, 259 responses were labeled as 'undecided' and excluded from the analysis. The remaining 49,693 responses were used for the analysis. More details are provided in the dataset readme file.

(a) (b)

Fig. 1: (a) Example of a *plain triplet comparison*. The question was "Which image looks more similar to the middle one?" (b) Simulation ground truth. The points simulate different distortion types (encoded by colors) and levels (encoded by point size) and are arranged in spirals. The reference level 0 is at the center, marked with a circle.

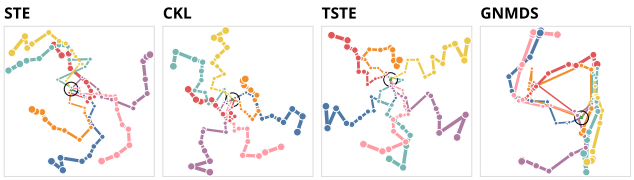

Fig. 2: 2D triplet embeddings for the spiral simulation data. The linear Pearson correlation between reconstruction and ground truth distances are from left to right $0.77, 0.76, 0.70, 0.54$. Normalized triplet errors are: 0.215, 0.212, 0.218, 0.229. Visual encoding as in Fig. 1b.

III. Multidimensional Triplet Embedding

A. Existing triplet embedding methods

Many triplet embedding methods have been proposed in the literature [26], e.g., GNMDS [17], STE & t-STE [18], CKL [19]. The general idea of the methods is to take the triplet comparison among entities as input and output their embeddings in an Euclidean space. The distances among points represent the entities' proximities based on the triplet comparisons. In this work, we rank the performance of several reconstruction methods using one simulation data with ground truth, using linear Pearson correlation coefficient r [27] between reconstruction and ground truth distances [28], see Fig. 3. From there, we observe the reconstruction of our dataset with the ranked methods.

B. Logistic triplet embedding (LTE) with regularization

In order to apply MLE, we need to define the model probability of the response $R =$ left to the triplet questions of which image is perceptually closest to the pivot. Let $x_l, x_p, x_r \in \mathbb{R}^n$ be the embedded points for the left, center, and right images of a triplet, respectively. Then we take the

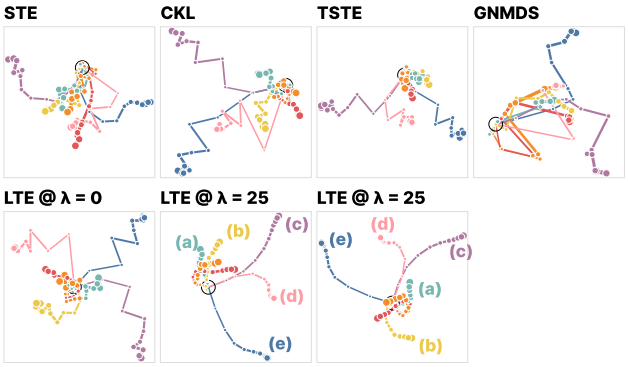

Fig. 3: 2D triplet embeddings of our image quality dataset. Color-coded distortion types: • highsharpen, • motionblur, • lensblur, • jpeg2000, • reference, • jitter, • colordiffusion, • multinoise. Top row: Classical methods with normalized triplet errors 0.139, 0.129, 0.126, 0.138 (left to right). Bottom row: The logistic triplet embedding (LTE) without regularization ($\lambda = 0$, left). On the right, with $\lambda = 25$ and two different initializations resulting in topologically different sequences of branches. Normalized triplet errors 0.127, 0.130, 0.128.

logistic function of the signed difference of the distances, $d = ||x_r - x_p|| - ||x_l - x_p||$, and set

$$P(R = \text{left}) = \frac{1}{1 + e^{-\alpha d}}, \tag{1}$$

we choose $\alpha = \log 3$, which scales the units in the embedding space according to just-noticeable differences (JND). For a difference $d = 1$, we obtain $P(R = \text{left}) = 0.75$, which corresponds to 1 JND. This also matches the design of our database in which the stimuli for each distortion type were chosen with a spacing of approximately 0.25 JNDs between consecutive distortion levels. This model is similar to that used in stochastic triplet embedding.

To smooth noisy embeddings, we introduce a form of Tikhonov regularization [29] to the MLE by adding a penalty term to its loss function. To this end, we use the sum of the squared finite difference approximations of the second derivatives taken at all coordinate points of the sequences of embedded images of each distortion type. The term is multiplied by a Lagrange factor λ that controls the degree of smoothing.

C. Results

We computed triplet embeddings for our image quality dataset and also from simulated triplets with 2D ground truth as shown in Fig. 1b. The simulation parallels the empirical data, simulating 7 distortion types at 12 levels each and using the same triplets and the same number of ratings as those in the dataset. The probabilities for the triplet responses were computed using Equation (1).

Only if the triplet embedding methods are able to recover the spiral ground truth shape from the simulation data, can we expect any meaningful reconstructions for the subjective study data. The results, shown in Fig. 2, confirm for the methods STE, CKL, TSTE, and GNMDS that they indeed are able to recover more or less the rough outline of the ground truth spiral shape.

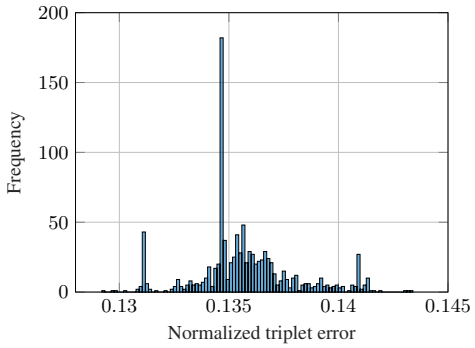

Fig. 4: The distributions of normalized triplet errors over 1000 initial configurations of embedding method STE for the subjective data of our image quality dataset.

After validating that the reconstructions work for a data set of the present format, we applied them to our experimental data set. The results for STE, CKL, TSTE, GNMDS, and LTE ($\lambda = 0$), shown in Fig. 3, reveal a few expected polyline structures for the different distortion types, albeit quite noisy. Thus, our dataset supports the hypothesis of their existence. When adding the Tikhonov regularization with $\lambda = 25$, these linear strands become very prominent, clearly demonstrating the potential of multidimensional image quality reconstruction.

Our data analysis shows the challenges of this approach. In triplet embedding, a loss or stress function such as the negative log-likelihood is minimized for the statistical model used in STE and LTE. Optimizers can get stuck in local minima, and this also happens in this application. We performed 1000 runs for STE with random initializations and recorded the normalized triplet loss for all generated solutions. These values are shown in the histogram in Fig. 4. Obviously, there are not just a few but many local optima, each suggesting a distinctly different geometric pattern for the reconstruction.

Thus, different global topologies may emerge for the same dataset using different reconstruction techniques or just different initializations in their optimizers. This is demonstrated with one example in the bottom row of Fig. 3 for the LTE method with regularization. We labeled the branches by their corresponding distortion types (color-coded) for the two results from different initializations. In clockwise order we have a-b-c-d-e for the first and a-b-e-d-c for the other case. Note that these symbol sequences cannot be aligned by an isometric transformation of the point configurations of the reconstructions.

IV. Conclusions

This work recalls the topic of multidimensional perceptual image quality assessment and shows its potential and challenges. We contributed an annotated dataset of distorted images with triplet comparisons. We also introduced a new triplet embedding technique adapted to perceptual quality scaling, and with a regularization term that helps to align sequences of increasingly distorted images in the embedding. We plan to integrate the triplet comparison information directly into their embeddings as a layout enrichment to get more faithful insight into the global topologies.

REFERENCES

[1] CCIR XVIIth Plenary Assembly, "Report 1082-1; studies toward the unification of picture assessment methodology," *Reports of the CCIR*, vol. 1, pp. 384–414, 1990.

[2] JS Goodman and DE Pearson, "Multidimensional scaling of multiply-impaired television pictures," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 6, pp. 353–356, 1979.

[3] Lena Linde, Hans Marmolin, and Sten Nyberg, "Visual effects of sampling in digital picture processing-a pilot experiment," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11, no. 3, pp. 201–207, 1981.

[4] Friedemann Köster and Sebastian Möller, "Analyzing perceptual dimensions of conversational speech quality," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 2041–2045.

[5] J.B. Martens, "Multidimensional modeling of image quality," *Proceedings of the IEEE*, vol. 90, no. 1, pp. 133–153, 2002.

[6] Boris Escalante-Ramírez, Jean-Bernard Martens, and Huib de Ridder, "Multidimensional characterization of the perceptual quality of noise-reduced computed tomography images," *Journal of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 317–334, 1995.

[7] Falk Schiffner and Sebastian Möller, "Defining the relevant perceptual quality space for video and video-telephony," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017, pp. 1–3.

[8] Robert C Streijl, Stefan Winkler, and David S Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[9] "Dimension-based subjective quality evaluation for video content," Recommendation ITU-T P.918, 2020.

[10] "Subjective diagnostic test method for conversational speech quality analysis," Recommendation ITU-T P.804, 2017.

[11] "Extension of ITU-T P.863 for multi-dimensional assessment of degradations in telephony speech signals up to fullband," Recommendation ITU-T P.863.2, 2022.

[12] LMJ Meesters and JBOS Martens, "Perceptual attributes of image quality in JPEG-coded images," *IPO Annual Progress Report*, vol. 32, pp. 147–153, 1997.

[13] Albert J Ahumada and Cynthia H Null, "Image quality: A multidimensional problem," *Digital Images and Human Vision*, pp. 141–148, 1993.

[14] Babak Naderi, Ross Cutler, and Nicolae-Cătălin Ristea, "Multidimensional speech quality assessment in crowdsourcing," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 696–700.

[15] Mason Macklem, "Multidimensional modelling of image fidelity measures," M.S. thesis, 2002.

[16] Vishwakumara Kayargadde and Jean-Bernard Martens, "Perceptual characterization of images degraded by blur and noise: experiments," *JOSA A*, vol. 13, no. 6, pp. 1166–1177, 1996.

[17] Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David Kriegman, and Serge Belongie, "Generalized non-metric multidimensional scaling," in *Artificial Intelligence and Statistics*. PMLR, 2007, pp. 11–18.

[18] Laurens van der Maaten and Kilian Weinberger, "Stochastic triplet embedding," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 2012, pp. 1–6.

[19] Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai, "Adaptively learning the crowd kernel," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 673–680.

[20] Hui Men, Hanhe Lin, Mohsen Jenadeleh, and Dietmar Saupe, "Subjective image quality assessment with boosted triplet comparisons," *IEEE Access*, vol. 9, pp. 138939–138975, 2021.

[21] Lina Jin, Joe Yuchieh Lin, Sudeng Hu, Haiqiang Wang, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo, "Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis," *Electronic Imaging*, vol. 2016, no. 13, pp. 1–9, 2016.

[22] Mohsen Jenadeleh, Johannes Zagermann, Harald Reiterer, Ulf-Dietrich Reips, Raouf Hamzaoui, and Dietmar Saupe, "Relaxed forced choice improves performance of visual quality assessment methods," in *15th International Conference on Quality of Multimedia Experience (QoMEX)*, 2023, pp. 37–42.

[23] Mohsen Jenadeleh, Raouf Hamzaoui, Ulf-Dietrich Reips, and Dietmar Saupe, "Crowdsourced estimation of collective just noticeable difference for compressed video with the flicker test and QUEST+," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–17, 2024, Early Access.

[24] Michela Testolina, Vlad Hosu, Mohsen Jenadeleh, Davi Lazzarotto, Dietmar Saupe, and Touradj Ebrahimi, "JPEG AIC-3 dataset: Towards defining the high quality to nearly visually lossless quality range," in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2023, pp. 55–60.

[25] Davi Lazzarotto, Michela Testolina, and Touradj Ebrahimi, "Subjective performance evaluation of bitrate allocation strategies for mpeg and jpeg pleno point cloud compression," *arXiv preprint arXiv:2402.04760*, 2024.

[26] Leena Chennuru Vankadara, Michael Lohaus, Siavash Haghiri, Faiz Ul Wahab, and Ulrike von Luxburg, "Insights into ordinal embedding algorithms: A systematic evaluation," *Journal of Machine Learning Research*, vol. 24, no. 191, pp. 1–83, 2023.

[27] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen, "Pearson correlation coefficient," *Noise Reduction in Speech Processing*, pp. 1–4, 2009.

[28] Jan De Leeuw and Patrick Mair, *Shepard diagram*, pp. 1–3, John Wiley & Sons, Ltd, 2015, Available online at: https://doi.org/10.1002/9781118445112.stat06268.pub2.

[29] Rainer Kress, *Tikhonov Regularization*, pp. 243–258, Springer Berlin Heidelberg, Berlin, Heidelberg, 1989.