Response to Affine Transforms of Image Distance Metrics and Humans

Paula Daudén-Oliver, Nuria Alabau-Bosque, Jorge Vila-Tomás, Jesús Malo, Valero Laparra
Universitat de València, Image Processing Laboratory, Valencia, 46022, Spain

Abstract

The standard approach to testing image quality models with deep architectures is through correlation with human opinion of distortions typically found in digital media. RAID-database presents a more human way of testing distorted images with affine transformations that are the ones found in nature. We have selected 6 image quality metrics (2 convenient references and 4 state-of-the-art) to test their alignment with human behavior with the same psychophysical method: Maximum Likelihood Difference Scaling. Although perceptual metrics are designed to predict human perception, we found that none of them accurately replicate human response curves for the three proposed affine transformations. Specifically, we analyzed the ranking regard human responses to different images within the same distortion, the ranking with regard human sensitivity of single images when different affine distortions are applied, and the shape of the MLDS response curve.

Keywords: Perceptual Metrics; Human Response; Affine Transformations

Introduction

Typically, subjective image quality models have been evaluated according to their ability to correlate with human opinion (Zhang et al., 2018) in databases containing a wide range of generic distortions (Ponomarenko et al., 2015). One limitation of assessing model performance in this way is leaving out other important phenomena within human vision (Martinez, Bertalmío, & Malo, 2019; Alabau-Bosque et al., 2024). RAID-database presents human responses obtained using the Maximum Likelihood Difference Scaling method (Maloney & Yang, 2003), a classical method in psychophysics. The transformations used in RAID-database are Gaussian noise and affine transformations, that are the ones that can be found in nature: rotation, translation and scaling. In this work, 5 reference images from RAID-database are used to test models and evaluate their performance compared to humans. This comparison can be made in terms of shape, response range and dependence on the reference images.

Methods

Five reference images from RAID-database and their human responses were selected to evaluate the sensitivity curves of 6 image quality models (RMSE, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2020), PerceptNet (Hepburn et al., 2020), and PIM (Bhardwaj et al., 2021)). As in RAID-database, the MLDS method was applied to the models obtaining response curves from 0 to 1. In humans, the curves are scaled with the internal noise obtained in the measurements. As models do not have internal noise, the curves were scaled to the distance between the reference image and the most distorted one for each image and model.

Results

Figure 1 shows the results obtained by the models and the human responses. On the top, the reference images are displayed with different colors according to the color curves. First row shows the human response from RAID-database and the other six rows are the response curves of the different models. Columns are the four transformations (Gaussian noise, rotation, translation, scaling). It is clear visually that none of the perceptual models (neither the RMSE) are aligned with human responses.Next, we provide numerical comparison comparing each model with the human responses in three different ways: the ordering inside each plot (intra-distortion), the ordering with regard other distortions (inter-distortion), and the shape of the curve.

Table 1 results for the ordering inside each plot. In particular, provides the Spearman correlation for the maximum values of the curves between each model and the human response for each distortion. No model correlates with humans in a statistically significant way. However, models perform better when dealing with Gaussian noise and Scaling, where all the correlations are positive. The best model is DIST, although the correlations for rotation and translation small.

Table 1: Intra-distortion evaluation: Spearman correlation of each models prediction with humans predictions for each distortion.

	All		Gaussian Noise		Rotation		Translation		Scaling	
MODELS	Corr	p-val	Corr	p-val	Corr	p-val	Corr	p-val	Corr	p-val
RMSE	-0,06	0,81	0,60	0,35	-0,90	0,08	0,30	0,68	0,30	0,68
SSIM	-0,09	0,70	0,80	0,13	-0,60	0,35	-0,50	0,45	0,50	0,45
LPIPS	0,10	0,66	0,80	0,13	-0,90	0,08	0,30	0,68	0,70	0,23
DISTS	0,28	0,23	0,70	0,23	-0,20	0,78	0,20	0,78	0,80	0,13
PerceptNet	-0,02	0,95	0,70	0,23	-0,70	0,23	0,30	0,68	0,30	0,68
PIM	0,25	0,29	0,80	0,13	-0,90	0,08	0,40	0,52	0,60	0,35

Results analyzing the behavior among distortions is presented in Table 2. It provides, the average correlations (ordering rank) for each image across all distortions between the models and the humans. Results indicate that there is no correlation between human responses and any of the models in this case. The best model is RMSE with an average correlation of 0.32, but a huge standard deviation of 0.41.

Finally, we analyze the difference in the curve shape. Table 3 shows the average absolute differences (L1 norm) between the curves shapes for models and humans. To remove the scaling effect, we normalized each curve by its maximum before computing the difference. In this case the models get a reasonable shape for the affine transformations (which are al-

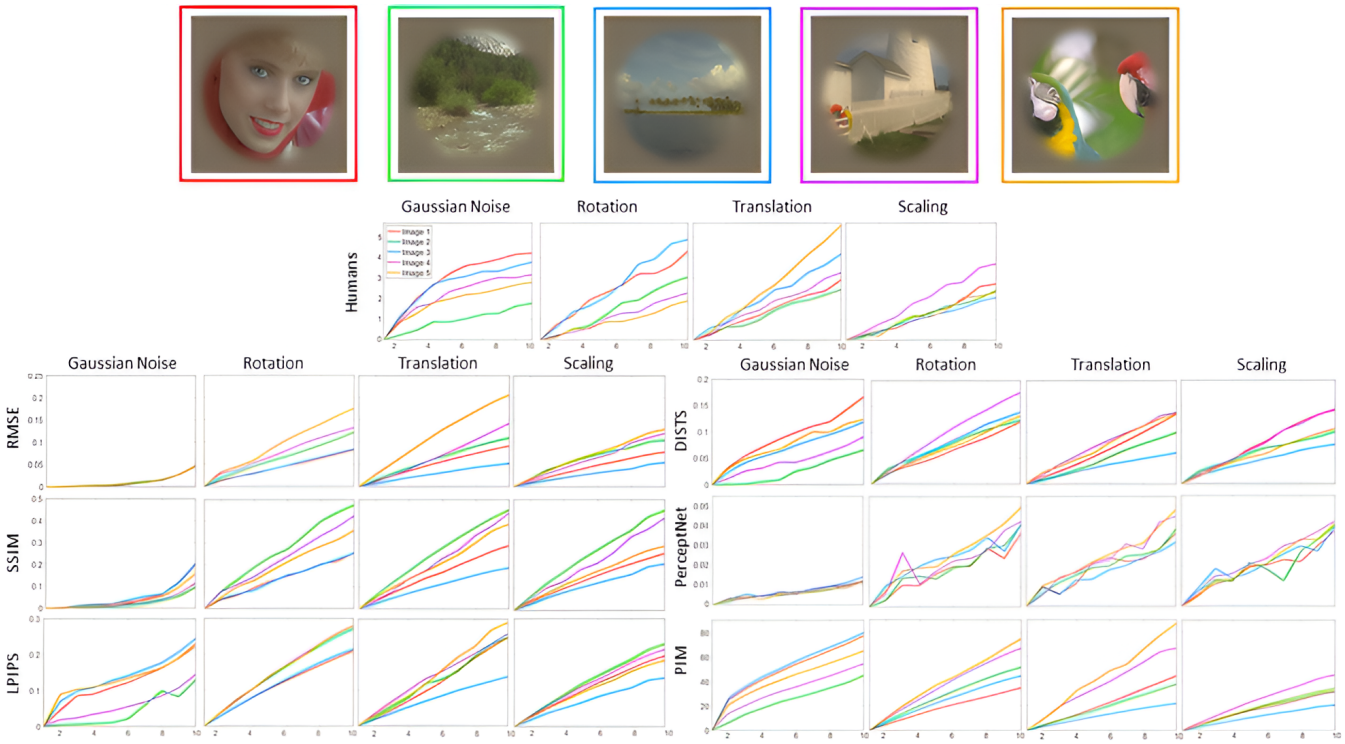

Figure 1: Tested images (top) and MLDS curves of the humans (top) and 6 models (rows) for the different distortions (columns).

Table 2: Inter-distortion evaluation: Average and standard deviation of the Spearman correlation between each model and human predictions for each image.

RMSE	SSIM	LPIPS	DISTS	PerceptNet	PIM
0,32 (0,41)	0,16 (0,57)	0,32 (0,50)	0,20 (0,49)	0,08 (0,52)	0,16 (0,54)

most linear), and the interesting part is in the Gaussian noise, where only PIM gets an error smaller than 10%.

A one-way ANOVA test was applied to determine whether the differences between groups were statistically significant (results not shown). Then a Tukey-Kramer post-hoc test was used to identify which specific pairs of groups showed statistically significant differences. Results indicate that for Gaussian noise, RMSE and SSIM exhibit larger differences from human responses than the other models in almost all images. This is in agreement with the fact that all the other metrics have better human behavior. Only PIM shows statistically significant differences in all images for this distortion. For rotation, PerceptNet shows higher differences than the other models only in image 3. For translation and rotation, no significant differences were found between the models.

Conclusions

The principal conclusion is that none of the models reproduce the behavior of the human curves properly for the affine distortions. All of them work well for Gaussian noise as expected.

When analyzed in detail we can see that the ordering between images within the same distortion has no strong correlation between models and humans, giving better correlations for *translation* and *scaling* (apart from Gaussian noise). While not a good model the best one in this test is DISTS.

Table 3: Shape evaluation: Average and standard deviation difference between the curves of each model and the human ones (transposed).

	Gaussian Noise	Rotation	Translation	Scaling
RMSE	0.39 (0.06)	0.05 (0.02)	0.05 (0.02)	0.06 (0.04)
SSIM	0.39 (0.06)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)
LPIPS	0.18 (0.07)	0.05 (0.02)	0.03 (0.01)	0.06 (0.02)
DISTS	0.14 (0.05)	0.05 (0.03)	0.04 (0.01)	0.07 (0.03)
PerceptNet	0.15 (0.06)	0.08 (0.02)	0.06 (0.02)	0.07 (0.02)
PIM	0.06 (0.02)	0.05 (0.02)	0.03 (0.01)	0.06 (0.02)

Regarding the differences across distortions, all models exhibit behavior that diverges significantly from that of humans. This finding is particularly important, as it highlights the poor alignment between so-called perceptual metrics and human perception when evaluated using natural distortions (i.e., affine transformations).

Finally, in terms of shape, all models show a linear increase in rotation, translation, and scaling that is the same for humans. There is no statistical difference between models. Besides, the saturating behavior in Gaussian noise in humans is not present in the models, being more different for RMSE and SSIM. In shape terms the model that behaves better is PIM.

Acknowledgements

This work was supported in part by MICIIN/FEDER/UE under Grant PID2020-118071GB-I00 and PDC2021-121522-C21, in part by Generalitat Valenciana under Projects GV/2021/074, CIPROM/2021/056 and CIAPOT/2021/9, and Grant CIACIF/2023/223. Some computer resources were provided by Artemisa, funded by the European Union ERDF and Comunitat Valenciana as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV).

References

Alabau-Bosque, N., et al. (2024). *Invariance of deep image quality metrics to affine transformations.* Retrieved from https://arxiv.org/abs/2407.17927

Bhardwaj, S., et al. (2021). *An unsupervised information-theoretic perceptual quality metric.*

Daudén-Oliver, P., et al. (2025). *Raid-database: human responses to affine image distortions.* Retrieved from https://arxiv.org/abs/2412.10211

Ding, K., et al. (2020). Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. Retrieved from https://doi.org/10.1109%2Ftpami.2020.3045810 doi: 10.1109/tpami.2020.3045810

Hepburn, A., et al. (2020, oct). Perceptnet: A human visual system inspired neural network for estimating perceptual distance. In *2020 IEEE international conference on image processing (ICIP).* IEEE. Retrieved from https://doi.org/10.1109%2Ficip40778.2020.9190691 doi: 10.1109/icip40778.2020.9190691

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *J. Vis.*, *3*(8). doi: 10.1167/3.8.5

Martinez, M., Bertalmío, M., & Malo, J. (2019). In praise of artifice reloaded: Caution with subjective image quality databases. *Front. Neurosci.*, *13*. doi: 10.3389/fnins.2019.00008

Ponomarenko, N., et al. (2015). Image database TID2013: Peculiarities, results and perspectives. *Signal Proc. Im. Comm.*, *30*, 57-77.

Wang, Z., et al. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600-612. doi: 10.1109/TIP.2003.819861

Zhang, R., et al. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF CVPR*, 586-595.