PieAPP: Perceptual Image-Error Assessment through Pairwise Preference

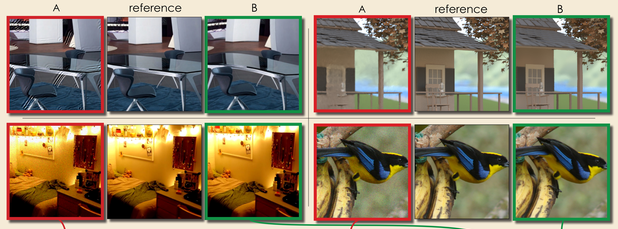
Ekta Prashnani* Hong Cai* Yasamin Mostofi Pradeep Sen

Motivation

Goal: Measure perceptual error between two images
Relevance: Full-reference quality assessment of images, objective functions for image restoration, image compression, Monte Carlo denoising, etc.
Challenges: Difficulty in modeling human perception of visual error, lack of large-scale dataset with noise-free labels or robust labeling schemes

Existing image-quality assessment approaches

"Which of the two distorted images, A or B, is visually closer to the reference?"



preferred by MAE/L1, RMSE/L2, SSIM, MS-SSIM, GMSD, VSI, PSNR-HMA, FSIMc, SFF, SCQI, DOG-SSIMc, Lukin et al., Kim et al., Bosse et al. (NR), Bosse et al. (FR)

preferred by humans and PieAPP

Limitations of hand-coded IQA models: difficulty in accurately modeling the complexity of the human visual system
Limitations of learning-based approaches: generalizability dependent on accuracy of ground-truth labels, size of dataset, etc.

Limitations of existing ground-truth labeling schemes

LIVE and CSIQ datasets: labeled with Mean Opinion Scores for image quality ➔ noisy and subjective

better image according to LIVE and CSIQ GT labels

TID datasets: labeled with Swiss tournament-based Mean Opinion Scores ➔ set dependent

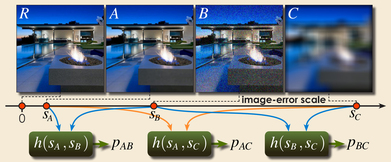
better image according to TID2008 GT labels

better image according to TID2013 GT labels

Contributions

- **Robust human response collection to capture the perception of visual error** that labels distorted image pairs with pairwise probabilistic human preferences — a strategy less prone to subjectivity or other inaccuracies of existing labeling schemes
- **Pairwise learning framework (PLF) to learn an image-error metric** that estimates image errors after training with human preference probabilities over distorted image pairs
- **Novel CNN-based perceptual image-error predictor** trained with the PLF framework, which demonstrates state-of-the-art performance on test images and distortion types not seen during training
- **Large-scale dataset** of distorted images and corresponding references labeled with the proposed human response collection scheme

Image Error from Pairwise Preference Probability



Assumption: There exists an underlying ordering of distorted images with respect to the corresponding reference image along a perceptual image-error scale.

Relating preference probability and image-error scores: Given the human preference probability, p_{AB}, of images A and B, we model the relation between image-error scores, s_A, s_B, and p_{AB} using the Bradley-Terry (BT) model:

$$p_{AB} = h(s_A, s_B) = \frac{1}{1 + e^{s_A - s_B}}$$

Learning to predict image-error scores: A function f with parameters θ that learns to predict the perceptual error between a distorted image and the reference can be trained using the BT model and human preferences on T image pairs:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{T} \sum_{i=1}^{T} \left\| h\left(f\left(A_i, R_i; \theta\right), f\left(B_i, R_i; \theta\right)\right) - p_{AB,i} \right\|_2^2$$

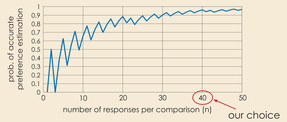
Human Response Collection

1. Query Amazon Mechanical Turk workers:

"Which of the two distorted images, A or B, is visually closer to the reference?"

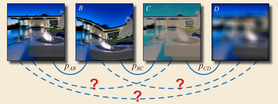
- results in triplets (A,R,B) and an associated label p_{AB} indicating probability of humans preferring A over B
- a subset of all possible pairs labeled in this manner

How many human responses needed per pairwise comparison to capture human probabilities reliably?



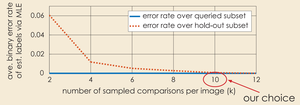
2. Estimate probabilities for unlabeled pairs:

Prob. labels for remaining pairs estimated using MLE



subset of pairwise preference prob. (e.g., p_{AB}, p_{BC}, p_{CD}) ➔ **maximum likelihood estimation** ➔ all pairwise preference prob. (e.g., $p_{AB}, p_{BC}, p_{CD}, p_{AC}, p_{BD}, p_{AD}$)

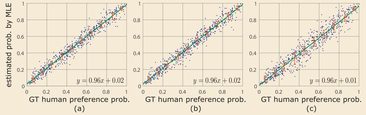
How many pairwise comparisons per image needed to accurately estimate the missing labels?



3. Reliability test of the prob. labels:

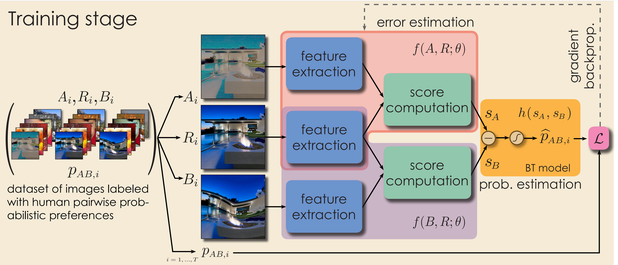
Are the obtained pairwise probabilistic preference labels accurately representing human pairwise preference?

(a) **BT model check:** estimated prob. obained by MLE on all possible GT pairwise prob. labels (100 human responses per pair)
(b) **Query reduction check:** estimated prob. obtained by MLE based on 10 comparisons per image with 100 human responses per pair
(c) **Overall reliability test:** estimated prob. obtained by MLE based on 10 comparisons per image with 40 responses per pair



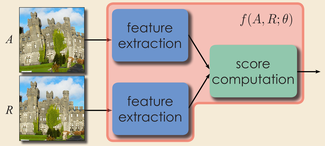
Pairwise Learning Framework

Training stage



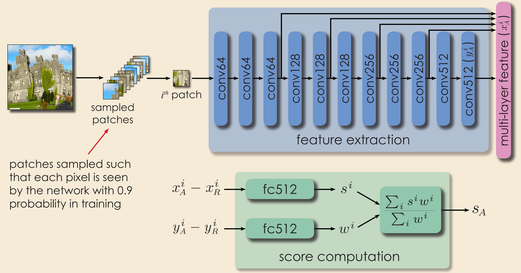
Key novelty: Training for probabilistic preference labels enables learning perceptual-error scores of images through the Bradley-Terry model, without the need for explicit ground-truth human labels for perceptual-error scores.

Evaluation stage



Interpretation: The pairwise differences of the predicted error scores indicate probability of preference of one image over another through the BT model.

Architectural choices



Training: Randomly-sampled patches are used to compute s_A. The estimated error scores for images A and B are used to compute the predicted probability which is trained to match the ground-truth preference probability.

Applying PLF-based training to other methods: The error-estimation function, f, in the PLF can be replaced with other learnable image-error estimators and trained on our pairwise preference data.

Results

Test set performance

METHOD	KRCC $\tilde{p}_{AB} \in [0,1]$	KRCC $p_{AB} \notin [0.35, 0.65]$	PLCC	SRCC
MAE	0.252	0.289	0.302	0.302
RMSE	0.289	0.339	0.324	0.351
SSIM	0.272	0.323	0.245	0.316
MS-SSIM	0.275	0.325	0.051	0.321
GMSD	0.250	0.291	0.242	0.297
VSI	0.337	0.395	0.344	0.393
PSNR-HMA	0.245	0.274	0.310	0.281
FSIMc	0.322	0.377	0.481	0.378
SFF	0.258	0.295	0.025	0.305
SCQI	0.303	0.364	0.267	0.360
DOG-SSIMc	0.263	0.320	0.417	0.464
Lukin et al.	0.290	0.396	0.496	0.386
Kim et al.	0.211	0.240	0.172	0.252
Bosse et al. (NR)	0.269	0.353	0.439	0.352
Bosse et al. (FR)	0.414	0.503	0.568	0.537
Our method (PieAPP)	**0.668**	**0.815**	**0.842**	**0.831**

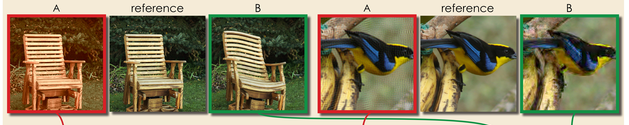
(see all references in the main paper)

- Test set distortions and reference images disjoint from the training set
- 31 test distortions types and 20 references
- 15 distorted copies per reference image
- Exhaustively labeled pairwise comparisons per reference to compute the correlation coefficients

preferred by MAE/L1, RMSE/L2, SSIM, MS-SSIM, GMSD, VSI, PSNR-HMA, FSIMc, SFF, SCQI, DOG-SSIMc, Lukin et al., Kim et al., Bosse et al. (NR), Bosse et al. (FR)

preferred by humans and PieAPP

Limitations



preferred by MAE/L1, RMSE/L2, SSIM, MS-SSIM, GMSD, VSI, PSNR-HMA, FSIMc, SFF, SCQI, DOG-SSIMc, Lukin et al., Kim et al., Bosse et al. (NR), Bosse et al. (FR), PieAPP

slightly preferred by humans (50%-60%)

PieAPP fails on close calls with unclear human preference.

Abbreviations:
- R: reference image
- A, B, C, D: distorted versions of the reference image
- GT: ground-truth
- BT: Bradley-Terry
- MLE: Maximum Likelihood Estimation
- NR: No-Reference
- FR: Full-Reference

Main References:
- Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," Signal Processing: Image Communication, 2015.
- Ponomarenko et al., "TID2008-A database for evaluation of full-reference visual quality assessment metrics," Advances of Modern Radioelectronics, 2009.
- Larson et al., "Most apparent distortion: Full-reference image quality assessment and the role of strategy," Journal of Electronic Imaging, 2010. (CSIQ dataset)
- Sheikh et al., "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Transactions on Image Processing, 2006. (LIVE dataset)
- Bradley et al., "Rank analysis of incomplete block designs: I. The method of paired comparisons," Biometrika, 1952. (Bradley-Terry model)
- Lukin et al., "Combining full-reference image visual quality metrics by neural network," SPIE/IS&T Electronic Imaging, 2015.
- Kim et al., "Deep learning of human visual sensitivity in image quality assessment framework," IEEE Conference on Computer Vision and Pattern Recognition, 2017.
- Bosse, et al., "Deep neural networks for no-reference and full-reference image quality assessment," IEEE Transactions on Image Processing, 2018.

email: {ekta,hcai,ymostofi,psen}@ece.ucsb.edu

website: http://cvc.ucsb.edu/graphics/Papers/CVPR2018_PieAPP/

source code: https://github.com/prashnani/PerceptualImageError

acknowledgements: NSF grants IIS-1321168 and IIS-1619376, and AI Grant (awarded to Ekta Prashnani)

*joint first authors