EVALUATING PERCEPTUAL DISTANCE MODELS BY FITTING BINOMIAL DISTRIBUTIONS TO TWO-ALTERNATIVE FORCED CHOICE DATA

Alex Hepburn, Raúl Santos-Rodríguez
School of Engineering Mathematics
Univ. of Bristol, United Kingdom

Javier Portilla
Instituto de Óptica,
C.S.I.C., Spain

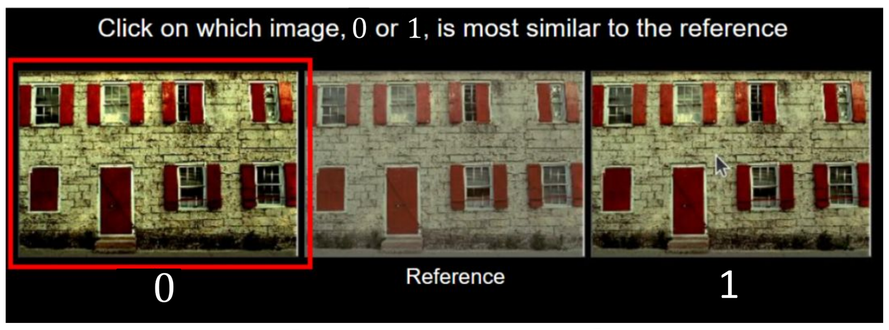
ABSTRACT

The Two Alternative Forced Choice (2AFC) paradigm in psychophysics (PF) is more simple and robust than the Mean Opinion Score (MOS). However, unlike MOS, it does not enable direct correlation between model predictions and PF data. In large datasets where images and distortions are combined randomly, deriving rankings from 2AFC PF data is not feasible. To address this problem researchers have trained neural networks to predict 2AFC PF responses from model distances. Instead, we set a robust and efficient distance-model evaluation based on binomial decisions in a purely probabilistic frame.

INTRODUCTION

- ***Two-Alternative Forced Choice (2AFC)*** *psychophysical (PF) experiment:*

  A triplet *is presented to the subjects:* $d_0 > d_1$?



- ***Goal:*** *to evaluate the fitness of distance models* $\{d_j(x, y)\}$ *to predict a (large) given set of 2AFC PF responses.*

- ***Problem:***
  2AFC requires shared images between triplets to yield image quality rankings (similar to MOS). Big PF databases, like ***BAPPS*** [1]*, don't share images between triplets => ranking unfeasible*

- ***Previous solution:***
  Train ad hoc neural networks (NNs) to predict the PF answers:
  - *For each triplet, the input is the pair of distances (d0,d1), and the NN is trained to predict the sum of the PF responses for that triplet.*

  Problems:
  - *PF responses are partially random, and the NN disregards this.*
  - *The NN does not model the decision process; it lacks "likelihood".*
  - *Each NN is only valid for a given distance model and* M *subjects.*
  - *Training each NN is costly in time and computation.*
  - *NNs are "black boxes" and lack interpretability.*

METHOD

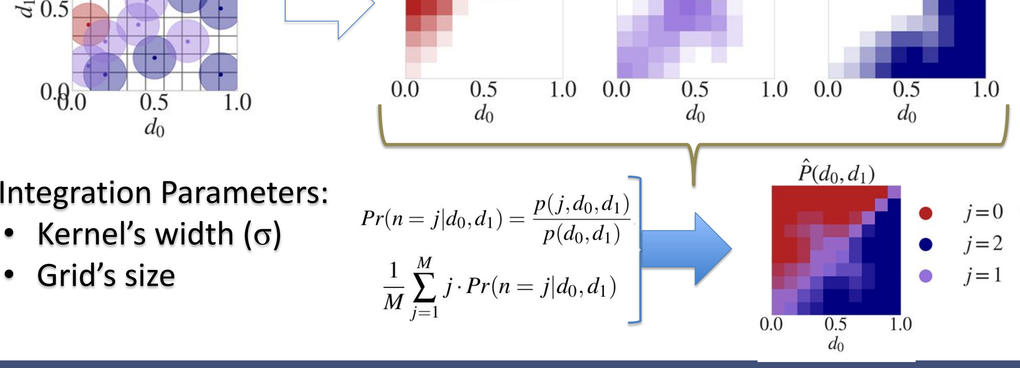
- *Model the PF data as* ***binomial decisions from*** **M** ***observers, given a distance model*** **d***:*
  - *1. Obtain an estimate of the density of* $p(j, d_0, d_1)$*, being* j *the number of people choosing the right image* x_1 *for each triplet, by computing these distances for each triplet and where the data cloud is smoothed using a Gaussian kernel.*
  - *2. Calculate* $p(d_0, d_1)$*, by adding the previous densities.*
  - *3. Obtaining the conditional probabilities of a given score* j *by:*

$$Pr(n=j|d_0,d_1)=\frac{p(j,d_0,d_1)}{p(d_0,d_1)}, j=1\dots M.$$

  - *4. The maximum likelihood parameter of a binomial choice running* M *times, given the conditional probabilities, is:*

$$\hat{P}(d_0,d_1)=\frac{1}{M}\sum_{j=1}^{M} j\cdot Pr(n=j|d_0,d_1)$$

- *Example, for* M = 2 *(*j *from* 0 *to* 2*):*



Integration Parameters:
- Kernel's width (σ)
- Grid's size

EXPERIMENTS

- ***Evaluation Metrics***

  AJ: *Comparing the most likely values coming from the binomial distributions* $\mathcal{B}(M,\hat{P}(d_0(t),d_1(t)))$ *with the actual judgments* n(t)

$$\mathrm{AJ}(\{n(t),t=1\dots T\},\hat{P},M)=100-\frac{100}{T}\sum_{t=1}^{T}\left|\frac{\lfloor(M+1)\hat{P}(d_0(t),d_1(t))\rfloor-n(t)}{M}\right|$$

  NLL: *Negative log-likelihood of the actual judgments according to the binomial model* $\mathcal{B}(M,\hat{P}(d_0(t),d_1(t)))$

$$\mathrm{NLL}(\{n(t),t=1\dots T\},\hat{P},M)=-\frac{1}{T}\sum_{t=0}^{T}\log Pr_{\mathcal{B}}(n(t);M,\hat{P}(d_0(t),d_1(t)))$$

  2AFC Score: *Agreement of the binary decisions* ($d_0 > d_1$?) *to the actual judgments, only based on the distance model* d

$$\hat{p}\in\{0,1\}$$

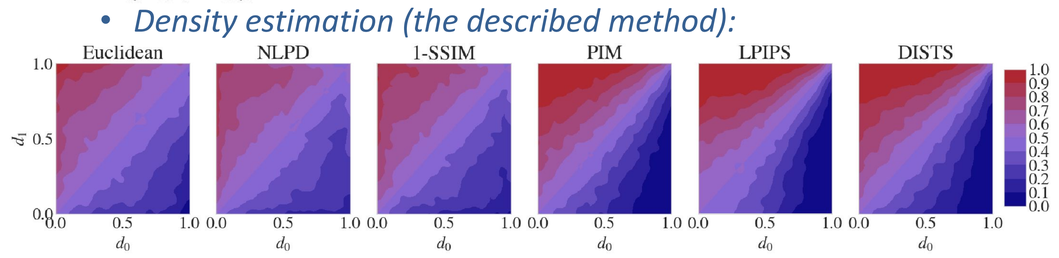
$$\mathrm{2AFC}=\frac{1}{T}\sum_{t=1}^{T}\left[\hat{p}(d_0(t),d_1(t))\cdot\frac{n(t)}{M}+(1-\frac{n(t)}{M})\cdot(1-\hat{p}(d_0(t),d_1(t)))\right]$$

- ***BAPPS dataset*** [1]***:*** 187k 64x64 *image patches arranged in triplets (reference, plus two distorted versions).*
  - *Training:* 151k *triplets,* M = 2, 302k judgements
  - *Validation:* 36k *triplets,* M = 5, 82k judgements
  - 425 *Distortions: noise, blur, color, compression, ANN-based, …*

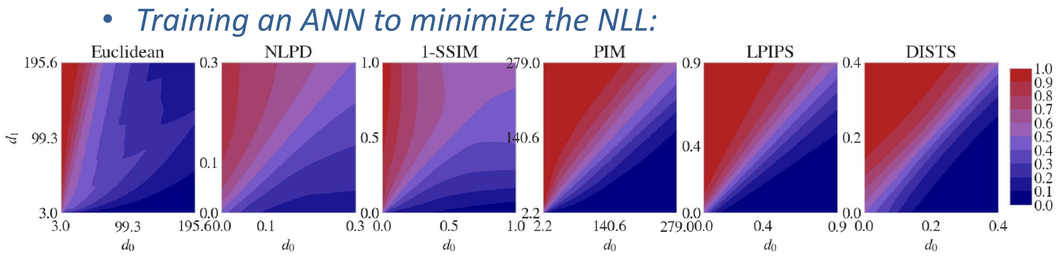
- ***Image distance models***
  - *Classical: Euclidean, NLPD* [2]*, SSIM*
  - *Deep-learning based: PIM* [3]*, LPIPS* [1]*, DISTS* [4]

RESULTS

- $\hat{P}(d_0,d_1)$ ***estimation***
  - *Density estimation (the described method):*



  - *Training an ANN to minimize the NLL:*



- ***Distance models evaluation***

	Measure	Euclidean	NLPD	SSIM	PIM	LPIPS	DISTS
Density Estimation	$\mathrm{AJ}(n,\hat{P},M)$ %↑	75.89 ± 0.00	75.82 ± 0.00	76.17 ± 0.00	82.12 ± 0.00	82.43 ± 0.00	81.34 ± 0.00
	$\mathrm{NLL}(n,\hat{P},M)$ ↓	1.86 ± 0.00	1.86 ± 0.00	1.84 ± 0.00	1.49 ± 0.00	1.46 ± 0.00	1.53 ± 0.00
	2AFC Score % ↑	63.46 ± 0.00	63.31 ± 0.00	63.65 ± 0.00	70.09 ± 0.00	68.94 ± 0.00	69.12 ± 0.00
Neural Network	$\mathrm{AJ}(n,\hat{P},M)$ %↑	75.76 ± 0.09	75.79 ± 0.08	76.18 ± 0.18	82.07 ± 0.13	82.55 ± 0.03	81.40 ± 0.03
	$\mathrm{NLL}(n,\hat{P},M)$ ↓	1.87 ± 0.00	1.86 ± 0.00	1.84 ± 0.01	1.50 ± 0.01	1.46 ± 0.00	1.53 ± 0.00
	2AFC Score % ↑	62.43 ± 1.31	63.78 ± 1.20	64.19 ± 1.37	70.31 ± 1.34	70.02 ± 1.04	68.81 ± 0.49

CONCLUSIONS

- ***Advantages*** *(compared to non-model-based ANNs)****:***
  Simplicity, interpretability, flexibility, and computational efficiency

- ***Requirements*** *(compared to ranking-based and MOS)****:***
  It requires a large database for reliable density estimations

1. R. Zhang, P. Isola, A. Efros, E. Shechtman, and O.Wang, "The unreasonable effectiveness of deep features as a perceptual metric," In Proceedings of the IEEE CVPR, pages 586–595, 2018.
2. V. Laparra, J. Ballé, A. Bernardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized Laplacian pyramid," Electronic Imaging, 2016(16):1–6. 2016.
3. S. Bhardwaj, I. Fischer, J. Ballé, and T. Chinen, "An unsupervised information-theoretic perceptual quality metric," Advances in Neural Information Processing Systems, 33:13–24, 2020.
4. Ding K, Ma K, Wang S, Simoncelli EP., "Image Quality Assessment: Unifying Structure and Texture Similarity," IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022 May;44(5):2567-2581.