What Should Be Balanced in a "Balanced" Face Recognition Dataset?

Haiyu Wu, Kevin W. Bowyer

BMVC 2023

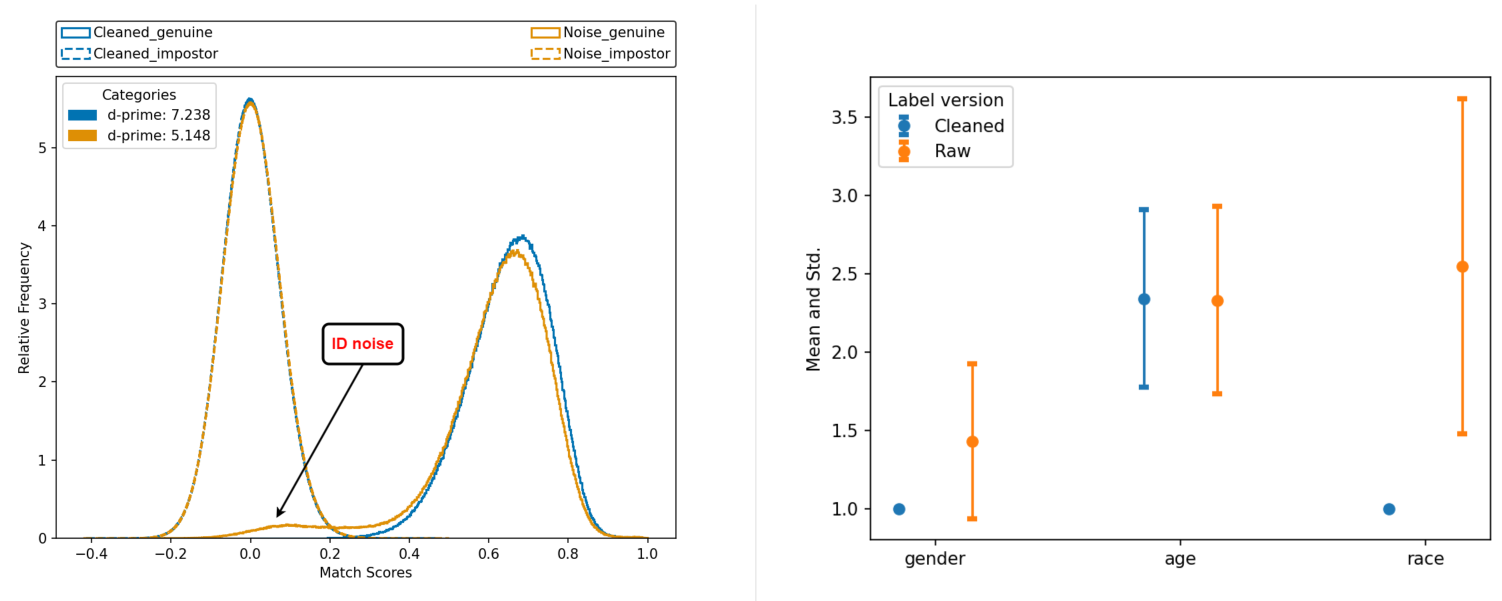
COMPUTER VISION @ ND

TL;DR

Problem: Existing "balanced" datasets are balanced on number of identities and number of images across demographics.
Solution: We show that, for face verification, the number of identities and the number of images across demographics in the test set do not drive accuracy differences.
Problem: Factors that are well-known to cause changes in accuracy are often not controlled in existing "balanced" datasets.
Solution: We assembled a Bias Aware test dataset, BA-test, that controls brightness, head pose, and image quality across demographics.
Problem: Existing accuracy disparity benchmarks may show accuracy disparity that is caused by non-protected attributes or protected attributes.
Solution: Our accuracy-disparity-focused benchmark controls the distribution of the non-protected attributes in order to ensure that an observed disparity is caused primarily by the protected attributes.

Identity De-noising and Protected Label Cleaning

- **Attribute classifier: FairFace**
- **Dataset: VGGFace2**



a) Identity de-noising b) Protected Label cleaning

Figure: a) Genuine / impostor distributions of random 200 **VGGFace2** identities before / after identity cleaning. b) Mean and s.d. of number of race, age, gender within each identity before / after cleaning.

Number of Identities / Images Doesn't Change Accuracy

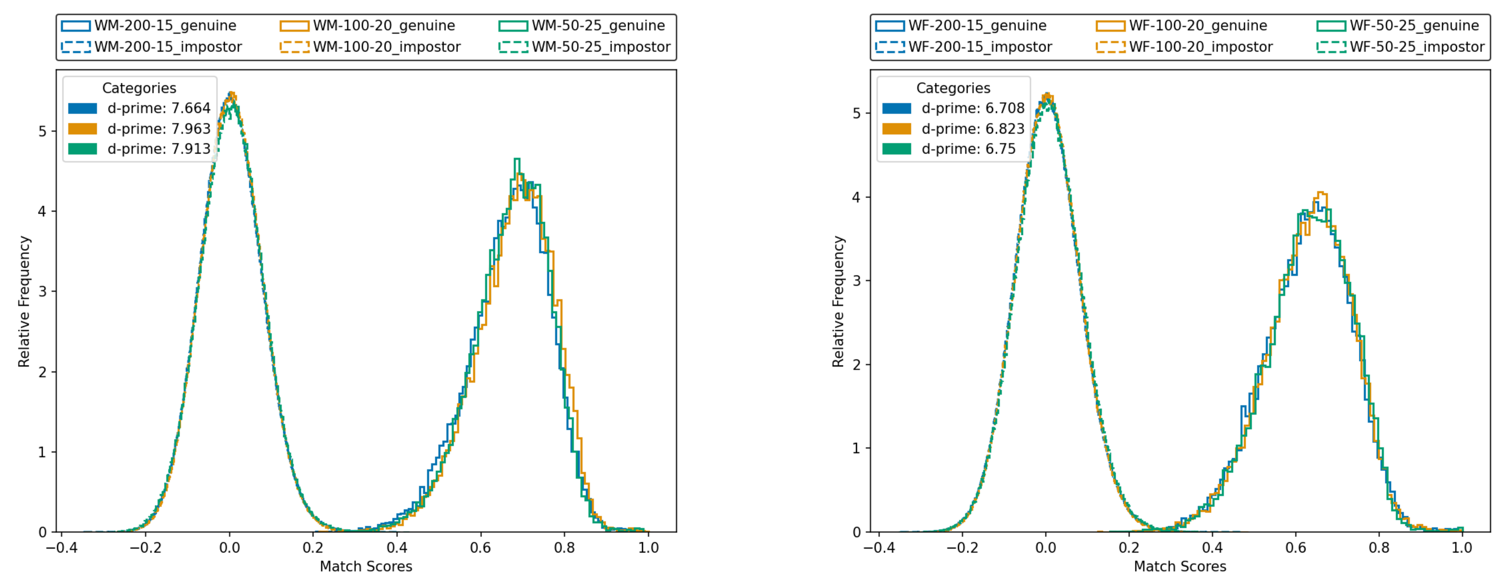
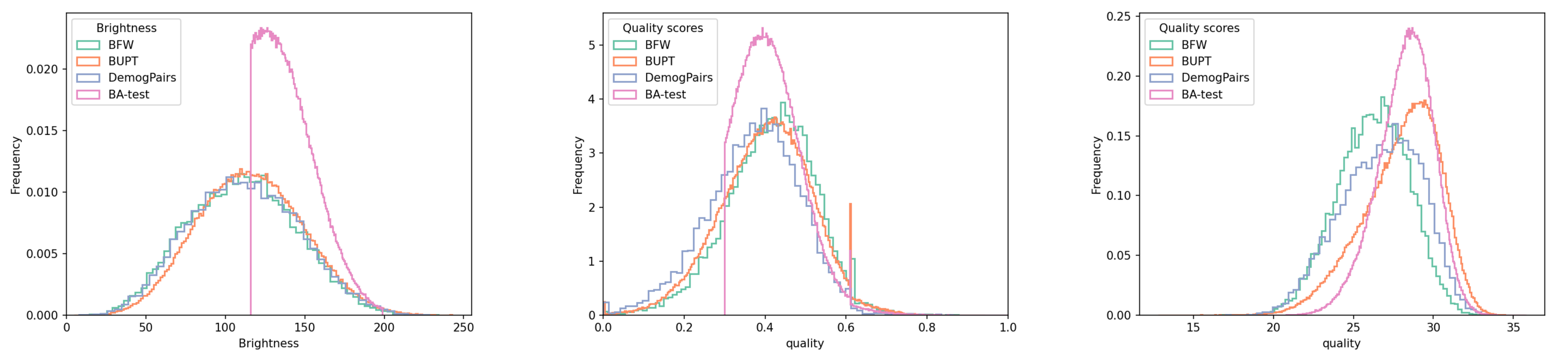


Figure: Distributions for varying number of identities and images per identity for White Male and White Female. "WF-200-15" means random 200 White Female identities with 15 images per identity.

Image Quality, Brightness, and Head Pose in Existing Balanced Datasets

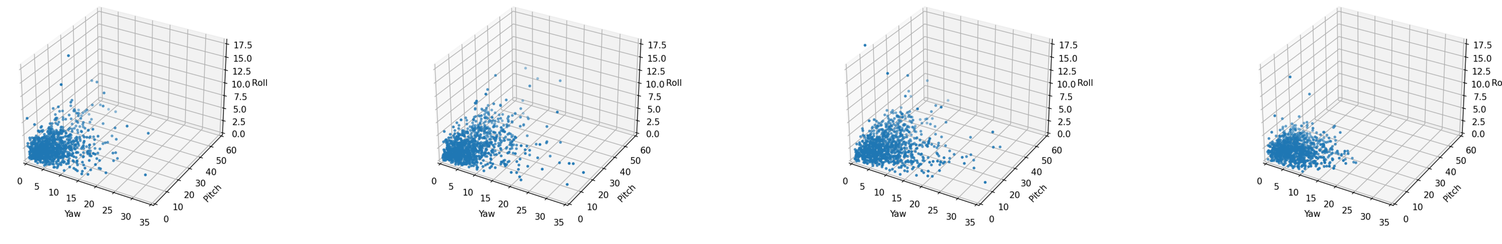
- **Datasets**: Balanced Face in the Wild (BFW), DemogPairs, BUPT-Balancedface, BA-test(ours)
- **Image quality**: MagFace, FaceQnet
- **Brightness**: Face Skin Brightness (FSB)
- **Head Pose**: img2pose



(a) Brightness (b) FaceQnet quality (c) MagFace quality

Figure: The brightness (a) and quality distributions (b), (c) of the existing balanced datasets.

The proposed dataset controls the brightness and image quality, but the others don't.



(a) BFW (b) BUPT-Balanedface (c) DemogPair (d) BA-test (**Ours**)

Figure: Head pose distributions of existing balanced datasets.

The proposed dataset has the most-frontal images, but the others don't.

Summary Statistics of Different Datasets

Statistic information						
Datasets	Data sources	IDs	Images	Subgroups	Age	ID denoise
DemogPairs	CWF, VGGFace, VGGFace2	600	10,800	6	✗	✗
BFW	VGGFace2	800	20,000	8	✗	✗
BUPT-Balancedface	MS-Celeb-1M	28,000	1.3M	4	✗	✗
RFW	MS-Celeb-1M, Face++ API	12,000	80,000	4	✗	✗
BA-test (ours)	VGGFace2	8,321	177,227	8	2	✓

Balanced factors						
Datasets	Head pose	Race	Quality	Brightness	ID	Gender
DemogPairs	✗	✓	✗	✗	✓	✓
BFW	✗	✓	✗	✗	✓	✓
BUPT-Balancedface	✗	✓	✗	✗	✓	✗
RFW	✗	✓	✗	✗	✗	✗
BA-test (ours)	✓	✗	✓	✓	✗	✗

Table: Existing demographically-balanced test datasets. Upper table gives source of data, number of identities, images, demographic groups, ages, and whether identity labels have been denoised. Bottom table shows factors balanced in each dataset.

Our dataset is good for understanding the cause of observed accuracy differences.

Benchmark on Demographics

Loss	Model	Train	AF	AM	diff.	WF	WM	diff.
MagFace	r50	Mv2	65.00	79.78	14.78	76.67	86.22	9.56
MagFace	r100	Mv2	81.56	94.44	12.89	89.44	96.56	7.11
ArcFace	r100	Mv2	81.56	93.11	11.56	90.11	97.11	7.00
ArcFace	r50	Glint	81.22	93.00	11.78	92.67	95.78	3.11
ArcFace	r100	Glint	90.00	96.78	6.78	95.78	98.78	3.00
Loss			BF	BM	diff.	IF	IM	diff.
MagFace	r50	Mv2	85.56	86.78	1.22	86.78	90.78	4.00
MagFace	r100	Mv2	91.00	94.22	3.22	96.00	96.11	0.11
ArcFace	r100	Mv2	91.56	94.11	2.56	94.56	95.56	1.00
ArcFace	r50	Glint	93.44	93.78	0.33	94.89	93.67	-1.22
ArcFace	r100	Glint	98.00	97.67	-0.33	98.56	97.22	-1.33

Table: True positive rates (%) with a false match rate of 10^{-5} and the best (green) and worst (red) accuracy for each face matcher across eight demographic groups. diff. is the highest TPR - the lowest TPR in each block. Mv2 and Glint are MS1MV2 and Glint360K.

The lowest accuracy is on Asian Females and the highest is on White Males.

Takeaways

- We demonstrate that datasets previously deemed "fair" or "balanced" for evaluation across demographics are not balanced on factors known to drive accuracy difference.
- We introduce the BA-test dataset, designed to support demographic accuracy disparity evaluations based on a better-balanced test set.
- We provide an accuracy-disparity-focused benchmark, revealing that current state-of-the-art models exhibit lowest accuracy on Asian females and highest on White males.

More About Haiyu Wu

- My webpage: https://haiyuwu.netlify.app/

Interesting Related Works

- **Brightness Affects Accuracy**: Face recognition accuracy across demographics: Shining a light into the problem (CVPRW 2023)
- **Must read if you use CelebA**: Consistency and Accuracy of CelebA Attribute Values (CVPRW 2023, best paper)
- **Facial hair dataset**: Logical Consistency and Greater Descriptive Power for Facial Hair Attribute Learning (CVPR 2023)
- **Facial hair effect**: Facial Hair Area in Face Recognition: Small Size, Big Effect