Departamento Inteligencia Artificial

https://pcr-upm.github.io/

Code available

Deconfounding demographic bias estimation in FER

Iván Ferre[1] , Roberto Valle[1], José M. Buenaposada[2] and Luis Baumela[1]

1. Universidad Politécnica de Madrid, 2. Universidad Rey Juan Carlos

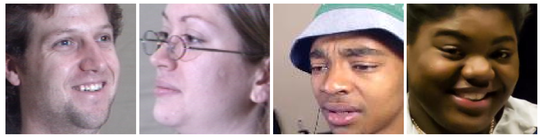
ABAW

CVPR JUNE 3-7, 2026 DENVER COLORADO

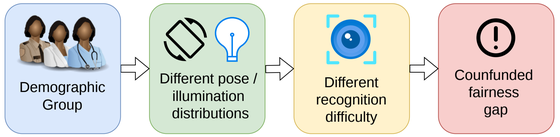
Problem

An unequal distribution of visual confounders such as head pose and illumination can bias fairness evaluation.

Why?



Smile Male | Neutral Female | Other Male | Happy Female



- FER fairness compares aggregate performance (*e.g.*, recall).
- Visual confounders (*e.g.*, head pose) affect FER performance.

Gender	Head Pose	
	Frontal	Profile
Female	69.01	68.72
Male	76.60	68.51
GAP	**7.59**	**-0.21**

Consequence

Measured GAP reflects test data acquisition conditions rather than intrinsic model bias.

Our Proposal

Use a **confounder-standardized evaluation** to compare demographic groups under the same visual conditions.

1. Predictions, labels, gender, confounder
2. Split into subgroup g and confounder bin b
3. Compute class-wise recall within each (g, b)
4. Average over classes
5. Average over bins
6. Compare groups (male / female)

Key equations

$M_g = \frac{1}{|\mathcal{C}|} \Sigma_{c \in \mathcal{C}} \text{TPR}_c^g$ Avg. recall across classes for group g (*e.g.*, $g \in \{\text{Female, Male}\}$)

$\text{GAP} = M_m - M_f$ Empirical gender fairness gap

$M_g^{std} = \frac{1}{|\mathcal{B}|} \Sigma_{b \in \mathcal{B}} M_{g,b}$ Standarized mean across bins b

$\text{GAP}^{std} = M_m^{std} - M_f^{std}$ Standardized gender fairness gap

$\Delta = \text{GAP} - \text{GAP}^{std}$ Confounding effect of the factor

$\Delta \approx 0$: estimate is stable
$\Delta > 0$: confounder inflates the gap
$\Delta < 0$: confounder masks the gap

Experimental setup

Controlled study: Multi-PIE

In-the-wild datasets: Aff-Wild2, AffectNet+, RAF-DB

Confounders: head pose (0°-15°, +15°) and illumination (0-125, 125-255)

FER model: ResNet-50 trained from scratch

Input images: Resized to 100 × 100 pixels

Controlled study in Multi-PIE results

Fairness gaps (%) under a pose confounding factor:

Subset	GAP	GAP^{std}	Δ
Balanced subset ($\approx$ pose/illum.)	4.51	4.60	-0.09
Biased subset ($\neq$ pose)	6.39	4.29	2.10

Introducing a gender-correlated head pose bias in the test set increases the measured fairness gap from 4.51% to 6.39%

In-the-wild analysis results

Fairness gaps (%) for pose and illum. confounding factors:

Database	Visual factor	GAP	GAP^{std}	Δ
Aff-Wild2	Head pose	4.23	6.25	-2.02
	Illumination		3.71	0.52
AffectNet+	Head pose	2.82	3.10	-0.28
	Illumination		2.46	0.36
RAF-DB	Head pose	-0.82	-1.74	0.92
	Illumination		-0.09	-0.73

Confounding effect is dataset dependent
Aff-Wild2: strongest pose effect
AffetNet+ and RAF-DB: less affected

Project PID2022-137581OB-I00 funded by MICIU/AEI/10.13039/501100011033 and FEDER, UE.

GOBIERNO DE ESPAÑA | MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES | Cofinanciado por la Unión Europea | AGENCIA ESTATAL DE INVESTIGACIÓN