LabellessFace

Fair Metric Learning for Face Recognition without Attribute Labels

Tetsushi Ohki[1,2], Yuya Sato[1], Masakatsu Nishigaki[1], Koichi Ito[3]
[1]Shizuoka University, Shizuoka, JP, [2]RIKEN AIP, Tokyo, JP, [3]Tohoku University, Miyagi, JP

IEEE International Joint Conference on Biometrics

web

Objective: Develop a novel face recognition algorithm that is fair across all demographic attributes, *even those not explicitly labeled*.

Our Idea: Instead of relying on demographic labels, treat each individual as a separate entity and *aim for fairness at the individual level*.

Contribution: **(i)** Propose ***class favoritism level*** which quantifies the degree of favoritism towards specific class across the entire dataset **(ii)** Propose ***fair class margin penalty*** to extend metric learning, enabling LabellessFace to improve fairness without target attribute labeling **(iii)** Comprehensive experiments have demonstrated that our proposed method is effective for enhancing fairness while maintaining authentication accuracy.

Motivation

Dependency to attribute labels

Traditional approaches to mitigating these biases heavily rely on demographic attributes.

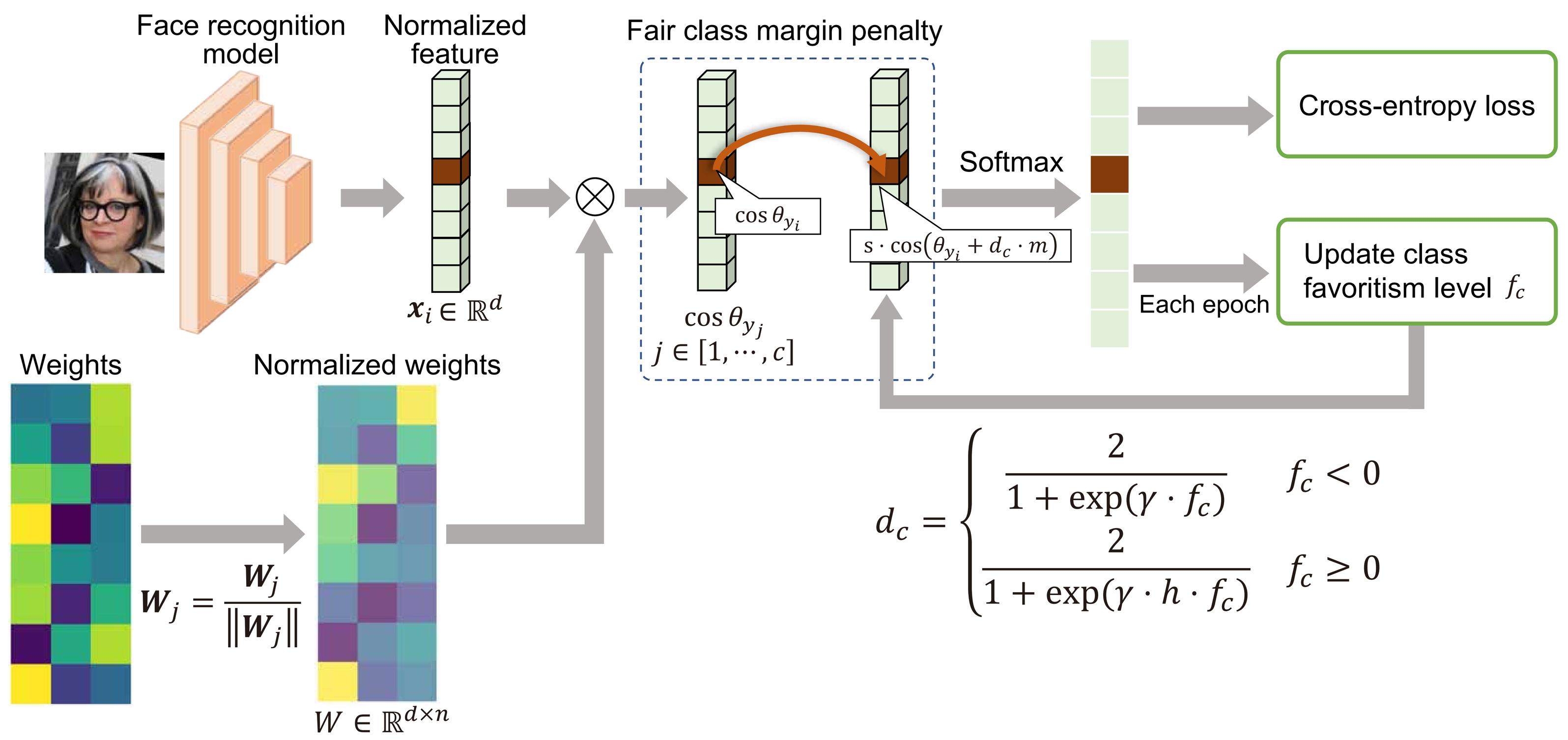
Scalability to large dataset

Creating large and fair datasets is costly in terms of recruiting participants and annotating attribute labels.

Can we improve a fairness notion without assuming the target attribute labels?

LabellessFace

Equalize authentication accuracy across individuals without assuming specific sensitive attributes, achieving fairness even for unknown attributes.



Fair Class Margin Penalty

A coefficient d_C (margin coefficient) is added to the basic ArcFace loss function to minimize the bias in individual authentication accuracy.

$$\mathcal{L} = -\log \frac{e^{s(\cos\theta_{y_i} + d_c \cdot m)}}{e^{s(\cos\theta_{y_i} + d_c \cdot m)} + \sum_{j=1, j \neq y_i}^{|C|} e^{s \cdot (\cos\theta_j)}}$$

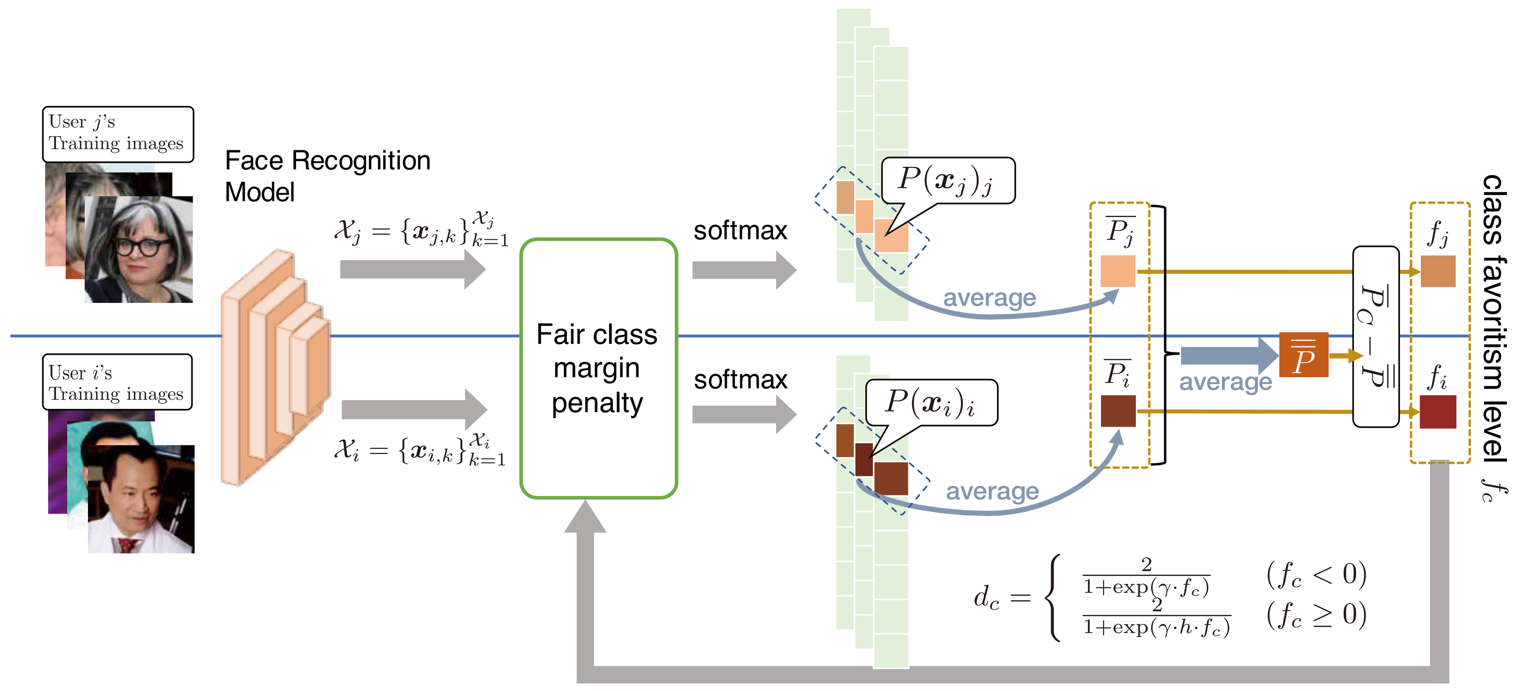
margin coefficient

$$d_c = \begin{cases} \frac{2}{1+\exp(\gamma \cdot f_c)} & (f_c < 0) \\ \frac{2}{1+\exp(\gamma \cdot h \cdot f_c)} & (f_c \geq 0) \end{cases}$$

class favoritism level

Class Favoritism Level

Class favoritism level quantifies the bias toward specific classes by measuring deviations in recognition accuracy compared to the overall class average.



Experiment

Dataset BUPT-Balanced face (training) / RFW and LFW (test)

Model ArcFace/MagFace/CIFP/MixFairFace/Proposed

Table: The performance and fairness evaluation results evaluated on LFW dataset. STD, Gini, SER were assessed when users were divided according to LFW 26 attributes.

	EER(↓)	AUC(↑)	STD(↓)	Gini(↓)	SER(↓)
ArcFace	0.09300	0.9665	0.01170	0.08292	2.766
MagFace	0.09867	0.9590	0.01127	0.08279	2.766
CIFP	0.09100	0.9614	0.01157	0.08845	3.038
Proposed	**0.09100**	**0.9681**	**0.01019**	**0.07398**	**2.525**

Our proposed method achieves consistently **high fairness across 26 attributes** with a single model

Takeaway

LabellessFace achieves **balanced accuracy across various attributes** by leveraging class favoritism levels and fair class margin penalties.

https://ohkilab.github.io/labellessface/ ohki@inf.shizuoka.ac.jp