顔認証システムの人種バイアスに影響を与える潜在的要因の調査

著者	佐藤 佑哉, 土屋 純, 成田 惇, 西垣 正勝, 大木 哲史
雑誌名	2022年暗号と情報セキュリティシンポジウム（SCIS2022）論文集
ページ	1-8
発行年	2022-01-12
出版者	電子情報通信学会
権利	(C)2022 The Institute of Electronics, Information and Communication Engineers
注記	2022年 暗号と情報セキュリティシンポジウム（SCIS2022) 開催形態 : オンサイト/オンライン ハイブリッド開催 オンサイト会場 : グランキューブ大阪（大阪府立国際会議場） 開催期間 : 2022年1月18日（火）～21日（金） セッション番号 : 4F1-4
著者版フラグ	publisher
URL	http://hdl.handle.net/10297/00029171

SCIS 2022 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

顔認証システムの人種バイアスに影響を与える潜在的要因の調査
Research on Potential Factors Affecting Racial Bias in Face Recognition

佐藤 佑哉 * 土屋 純 † 成田 惇 † 西垣 正勝 † 大木 哲史 †
Yuya Sato Jun Tsuchiya Jun Narita Masakatsu Nishigaki Tetsushi Ohki

あらまし 顔認証システムは非接触で対象の動作を必要とせず，使い勝手の良い生体認証であることから市場が広がっているが，多くの顔認証システムは人種間の認証精度に偏りが認められており，犯罪捜査に用いられた顔認証システムの誤認識を原因として有色人種の誤認逮捕が起きてしまうなど，人種差別に発展する問題が発生している．これまで，認証精度の偏りは，学習データにおけるセンシティブ属性の割合に起因すると考えられ，それらを軽減するための方法として，データセットにおける偏りの除去や，偏りを考慮したスコア正規化アルゴリズムの提案などが行われてきた．一方，認証システム全体に着目すれば，これらの対策を行った場合においても，取り除けない潜在的なバイアス要因が残存している可能性がある．本研究では，潜在的なセンシティブ情報の一調査として，顔画像の解像度，明度，彩度等の環境要因の変化が人種間の認証精度の偏りに与える影響の調査を行う．BFW データセット等の人種割合の偏りがないデータセットを用いた場合においても，テストセットの環境要因を変化させることで人種間の認証精度に偏りが生じるか検証する．

キーワード 顔認証，公平性，生体認証，人種バイアス

1 はじめに

人種や性別等の属性によって人工知能の誤検知率に違いが見られるなど，人工知能の公平性について問題視する声があがっている [1]．顔認証システムは非接触で対象の動作を必要とせず，カメラ等の安価な装置で実現できるといった利点があるなど，利便性の高い生体認証であることから市場が広がっている．しかし，顔認証システムにおいても，学習に用いる顔画像データセットの人種割合の偏りに起因する人種間の認証精度の偏りが指摘されている [2]．

顔認証システムは米国の法執行機関に対して提供され，犯罪捜査等の重要な場面で用いられてきたが，これらのシステムにおいても誤認識を原因とした有色人種の誤認逮捕などの事件が実際に発生している．これらの事件は顔認証システム利用の危険性として近年大きく取り上げられ [3]，大手 IT 企業による法執行機関への提供は停止され，それに伴い多くの企業が顔認証システムの利用停止やソフトウェアの販売停止といった措置へと発展した．このような社会問題を背景に，顔認証システムの公平性に関する検討の必要性が高まっていると言える．

人工知能や生体認証の利用において，利用者の公平性を保つ観点から，排除すべき利用者の属性を一般にセンシティブ属性という．米国の消費者信用機会均等法（The Equal Credit Opportunity Act, ECOA）では，人種や肌の色はセンシティブ属性として指定されており [4]，顔認証システムにおいては，人種や肌の色等の重要なセンシティブ属性が認証精度に影響を与えないように設計しなければならない．

顔認証モデルを学習させる際に一般的に使用される大規模顔画像データセット CASIA-WebFace [5]，VGGFace2 [6]，MSCeleb-1M [7] などは，Web サイトのスクレイピングによって無作為に収集・構築されている．しかし，現実には世界人口の 44%を占めるアジア人とインド人がデータセットには 8%ほどしか含まれないなど，データセットの人種割合が実際の人口割合とは大きく異なっている．これらの人種割合が偏ったデータセットを用いて学習された顔認証モデルは多数人種と比較して少数人種の認証精度が低下することが報告されている [8]．

顔認証システムにおいて，学習データセットや利用時に変動する環境要因が認証精度に影響を与えることを考

* 静岡大学情報学部
静岡県浜松市中区城北 3 丁目 5-1, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka, Japan
† 静岡大学大学院総合科学技術研究科
静岡県浜松市中区城北 3 丁目 5-1, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka, Japan

慮した認証方式は既に多く提案されており [9–11]，その中でも学習データセットの人種割合が顔認証システムの公平性に与える影響に関する検討が多く行われている．たとえば，BUPT-Balancedface 等の人種割合を統一した顔画像データセットで顔認証モデルの学習を行うことで，人種間の認証精度の偏りが軽減できることが示されている [10]．

一方，顔認証システムにおいて認証精度に影響を与える要因はデータセットだけでなく，たとえば照明条件といった登録-照合時の環境条件の違いもまた大きな要因と考えられる．これらに対して頑健であることは，顔認証システムの重要な評価項目である．公平性の評価においても，環境要因といったデータセット以外の要因に起因する認証精度の低下が属性間で異なる可能性が考えられる．しかし，このような顔画像を撮影する際に変動する環境要因が公平性に与える影響に関する検討は現時点で十分に行われていない．環境要因が公平性に影響を与えている場合，公平に学習を行った顔認証システムにおいても，顔画像を撮影する際の環境要因によって人種バイアスが顕在化し，公平性が損なわれる可能性がある．

本稿では，データセットの人種割合の偏り以外で人種間の認証精度の偏りに影響を与える可能性のある要因を潜在的要因として扱い，これらの要因が公平性に対して与える影響を評価する方法を提案する．またこれにより，公平性に影響を与えうる潜在的要因が存在することを示す．なお，公平性が示す意味については多くの議論があるが，本稿では，文献 [12] に基づき公平性を照合・識別両者を含む属性間の認証精度の偏りとして扱う．

表現学習に基づく顔認証システムを対象として，明度，彩度，コントラスト，解像度等の顔認証システムの利用時に起きうる環境要因の変動を加えたテストセットを用いて，これらの変動が公平性に与える影響の評価を行う．また，公平性評価指標のひとつであるデモグラフィックパリティ，等価オッズを拡張することで，環境要因による変動量を評価する指標 ΔΔ デモグラフィックパリティと ΔΔ 等価オッズを定義し，これらを用いて環境要因の変動が公平性に与える影響を定量的に評価する．これにより，実利用時の環境要因の変動が顔認証システムの人種バイアスに与える影響を明らかにする．

本研究の貢献は以下の 3 点である．

- 公平性定義のデモグラフィックパリティ，等価オッズの環境要因の変動に対する差分を評価することで，環境要因の変動に対して変化する人種間の公平性を評価する公平性指標を提案した．
- テストセットの環境要因を変動させて人種間の公平性を評価することで，実利用時に起きうる潜在的人種バイアスの可能性を示した．
- 人種割合のみを考慮したテストセットにおいては，実利用時の潜在的な人種バイアスまでを考慮した正しい公平性評価が行えない可能性を示した．

2 関連研究

2.1 顔認証システムの人種バイアス

顔認識システムは監視カメラの活用に伴い犯罪捜査等の重要な場面で用いられることもあり，人種に対する公平性の重要性は高く，多くの研究が行われている．

Buolamwini ら [2] は，Microsoft，IBM，Face++等が構築した顔認識システムは白人男性の誤認識率が 1%未満であるのに対して，黒人女性の誤認識率は 35%ほどであったことを報告している．

米国国立標準技術研究所（US National Institute of Standards and Technology, NIST）は，189 の顔認証ソフトウェアを対象に人種に対する公平性の調査を行い，1:1 認証において白人と黒人の誤検出率に 10 倍から 100 倍ほどの差が存在していたことや，1:N 認証において黒人女性の誤検知率が高いことを示した [13]．

Garvie ら [14] は，学習データの人種割合の偏りが人種バイアスに大きな影響を与えていることを指摘し，エンジニアは白人の割合が高いことから意図せずに人間による差別が介入している可能性や，肌の色がコントラストに影響を与えること，女性の化粧が認証精度に影響を与える可能性があることを指摘した．

2.2 データセットに関するアプローチ

機械学習の学習過程において，データセットの収集には人の手が加わるため，無意識のうちにバイアスが発生してしまう場合がある．そのため，顔認証システムにおいても人種バイアスの問題で最も着目されたのは学習データセットの人種割合である．

Wang ら [10] は，人種割合のバランスが取れた BUPT-Balancedface データセットを作成し，学習を行うことで従来の人種割合が偏ったデータセットで学習を行った場合と比較して人種バイアスが軽減されることを示した．

Faisal ら [15] は，人種間の認識精度の偏りが小さくなるようにデータオブジェクトの置換を繰り返すリサンプリングを行うことで，人種バイアスを引き起こすデータをデータセットから取り除く手法を提案した．また，学習データの収集において，DataAugmentation 等の手法も活用されることがあるが，Niharika ら [16] は GAN による DataAugmentation によって人種バイアスを軽減させる際の性能的な限界を指摘している．

2.3 モデルの構造に関するアプローチ

人種間の公平性を考慮した大規模な顔画像データセットを用意することは容易ではないため，人種に偏りのあ

る既存のデータセットにおいても公平に学習できるように，モデルの構造を工夫することで人種バイアス軽減に取り組む研究も行われてきた．

Philipp ら [9] は認証対象の人種を事前情報とし，人種間のスコアを正規化するアプローチを用いている．また，Wang らは Arcface 等の損失関数に含まれるハイパーパラメータを人種に応じて動的に変化させることで，人種割合が偏ったデータセットを用いて学習を行った場合でも人種間の認証精度の偏りを小さくできることを示した [10]．しかし，ハイパーパラメータを変化させる手法では人種間の認証精度の偏りは小さくなる一方で，多数人種は認証精度が低下してしまうという問題点がある．

3 調査手法

3.1 公平性の評価

機械学習の公平性については，デモグラフィックパリティ [12]，等価オッズ [17] という定義が広く用いられている．そのため，本稿においてもこの 2 つの定義を用いて公平性の評価を行う．それぞれの定義についての詳細を以下に示す．

3.1.1 デモグラフィックパリティ

デモグラフィックパリティはセンシティブ属性に依らず予測ラベルの比率が一定であるか否かによって属性間の公平性を評価する指標であり，予測ラベル $\hat{Y}$，センシティブ属性 $A \in \{0,1\}$，目的変数 $\hat{y} \in \{0,1\}$ に対して次式で定義される．

$$P\{\hat{Y}=\hat{y}|A=0\} = P\{\hat{Y}=\hat{y}|A=1\} \tag{1}$$

また，式 (1) より属性間の公平性を評価するデモグラフィックパリティ差 Δd は次のように示される．

$$\Delta d = \left| P\{\hat{Y}=\hat{y}|A=0\} - P\{\hat{Y}=\hat{y}|A=1\} \right| \tag{2}$$

$\Delta d = 0$ の時にデモグラフィックパリティを満たす．そのため，0 に近い値を取るほど属性間が公平であると考えられる．また，同様にして，デモグラフィックパリティ比を次式で定義できる．

$$r = \left| \frac{P\{\hat{Y}=\hat{y}|A=0\}}{P\{\hat{Y}=\hat{y}|A=1\}} \right| \tag{3}$$

生体認証は他人受入率が 10^{-6} など非常に小さい確率であることが多く，このため限られた試行回数において他人受入が発生しない，一度の他人受入が他人受入率に大きな影響を与えるといったことがある．このような場合，デモグラフィックパリティ比 r の変動もまた大きくなることが懸念される．これらを考慮し，本稿では，デモグラフィックパリティ比よりもデモグラフィックパリティ差を用いた評価を用いることを提案する．以降，デモグラフィックパリティ差を Δ デモグラフィックパリティと表記する．

3.1.2 等価オッズ

等価オッズはセンシティブ属性に依らず本人拒否率と他人受入率の割合が一定であるか否かによって公平性を評価する指標であり，次式で定義される．

$$P\{\hat{Y}=1|A=0,Y=\hat{y}\} = P\{\hat{Y}=1|A=1,Y=\hat{y}\} \tag{4}$$

式 (4) より，属性間の公平性を評価する等価オッズ差 Δe を次のように示すことができる．Δ デモグラフィックパリティと同様に $\Delta e = 0$ の時に等価オッズを満たし，0 に近い値を取るほど属性間が公平であると考えられる．

$$\Delta e = \left| \frac{P\{\hat{Y}=1|A=0,Y=0\}}{P\{\hat{Y}=1|A=0,Y=1\}} - \frac{P\{\hat{Y}=1|A=1,Y=0\}}{P\{\hat{Y}=1|A=1,Y=1\}} \right| \tag{5}$$

3.2 公平性変動の評価

本稿では，環境要因の変動に応じて潜在的な人種バイアスが顕在化するか確認するため，環境要因を変動させていない状態で人種間の認証精度に差が生じている場合にも評価が可能な公平性評価指標を定義する．

3.2.1 ΔΔデモグラフィックパリティ

環境要因を i 段階変動させた時の Δd を Δd_i とする．環境要因の変動に応じて潜在的な人種バイアスが顕在化する場合，$\{\exists i|\Delta d_i > \Delta d_0, i \neq 0\}$ となることが予想される．このような Δd_0 と Δd_i の間で生じる Δ デモグラフィックパリティの差分を評価するための指標 ΔΔデモグラフィックパリティ を次のように定義する．

$$\Delta\Delta d_i = |\Delta d_i - \Delta d_0| \tag{6}$$

式 (6) から自明なように，$\Delta d_0 = 0$ の時，$\Delta\Delta d$ はΔd に一致する．一方，$\Delta d_0 \neq 0$ の場合は，$\Delta\Delta d$ を用いて環境要因を変動させていない状態の Δ デモグラフィックパリティを考慮することで，環境変動による影響をより明確に分析できる．

3.2.2 ΔΔ等価オッズ

ΔΔデモグラフィックパリティ と同様に Δ 等価オッズの差分を評価するための指標 ΔΔ等価オッズ を次のように定義する．

$$\Delta\Delta e_i = |\Delta e_i - \Delta e_0| \tag{7}$$

ΔΔ デモグラフィックパリティや ΔΔ 等価オッズが大きな値を取るほど，環境要因を変動させていない状態と比

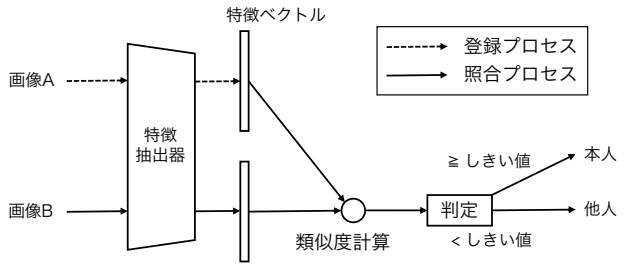


図 1: 1:1 認証の概要

較して潜在的人種バイアスが顕在化していると言える．本稿では，これらの評価指標で人種バイアスへの影響度を評価する．

3.3 認証シナリオ

本稿では，認証シナリオとして実利用に多く用いられる表現学習を用いた 1:1 認証を対象とする．1:1 認証は利用者の生体情報と，予め登録されている利用者の生体情報テンプレートの比較を行い，生体情報が利用者本人であるか否かを返答するシステムであり，スマートフォンのロック解除等の実利用に多く用いられる方式である．

本稿における 1:1 認証の概要を図 1 に示す．図 1 における顔画像ペア A，B は同時に入力されることもあるが，一般には A，B いずれかの顔画像から事前に特徴ベクトルを抽出し，テンプレートとして保管した上で，照合時に入力されたもう一方の特徴ベクトルとテンプレート間の類似度を算出して照合を行うことが多い．

特徴抽出過程では事前に学習されたモデルに顔画像を入力することで，顔特徴ベクトルを抽出する．特徴抽出アルゴリズムとしては，特に近年のニューラルネットに基づく方式では，多クラス分類器として学習した顔識別モデルの，最終層手前の層の出力などを用いることが多い．

その後，ふたつの顔特徴ベクトルに対してコサイン類似度等の類似度計算を行うことで，顔画像ペア A，B のスコアを算出する．この時，画像ペアが本人であるか否かを判定するしきい値は，実利用時にはあらかじめ検証用のデータを用いて利便性要件やセキュリティ要件を考慮して，誤非合致率や誤合致率の期待値が特定の値以下となるように設定されるが，本稿においては本人・他人画像ペアの全てのスコアを算出した後に，誤非合致率と誤合致率の割合が等しくなるようにしきい値を決定する．しきい値を決定した後，算出された各スコアがしきい値以上である場合に本人と予測し，しきい値未満であれば他人と予測する．

表 1: 実験環境

使用言語	Python3.8.10
GPU	GeforceRTX3090 × 3
CUDA コア数	10496 コア / GPU
CPU	AMD EPYC 7262
搭載メモリ	128GB
カーネルバージョン	5.4.0-58-generic
ディストリビューション	Ubuntu20.04.1LTS

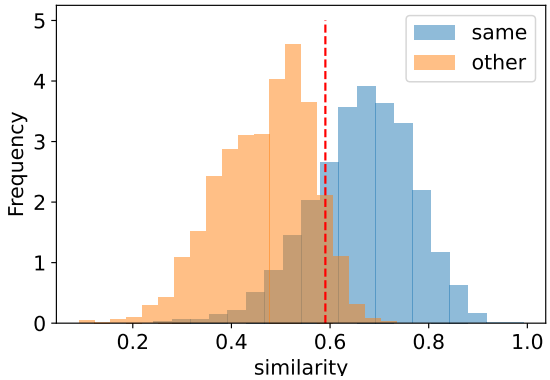


図 2: 本人・他人スコア分布の一例

4 実験

4.1 実験の手順

実験環境を表 1 に示す．

本研究では一般的に使用されている学習済みモデルである VGGFace [18] の入力層から全結合層までを特徴抽出器として使用する．顔画像ペアを特徴抽出器に入力し，得られたそれぞれの顔特徴ベクトルのコサイン類似度を算出する．

顔画像ペアは人種間の公平性を検討するために公開されているテストセットである BFW データセット [19] を使用する．BFW データセットは，Asian, Black, Indian, White の 4 人種各 200 名について，一名当たり 25 枚の顔画像から構成されるデータセットである．顔画像ペアは同じ人種から選択し，本人ペアおよび他人ペア各 6000 組の選択パターンを変更しつつ，10 分割交差検証により評価対象モデルの公平性について検証を行う．

10 分割交差検証のうちのひとつの検証における本人・他人の顔画像ペア各 6000 組をコサイン類似度により算出したスコアのヒストグラム図 2 に示す．しきい値は全ての組のスコアを算出した後に決定されたものであり，しきい値は赤波線に示した箇所となり，認証精度はしきい値の決定後に算出している．

4.2 環境要因の変動

本稿では，人種バイアスに影響を与える潜在的要因の一例として環境要因を調査対象とする．

人種特有の肌の色や明るさの違いに起因した潜在的人種バイアス要因が存在する場合，明度や彩度の変動によって影響を受ける度合いが人種間で異なる可能性があり，また，人種特有の彫りの深さや影の陰影の違いに起因した潜在的人種バイアス要因が存在する場合，コントラストや解像度の変動によって影響を受ける度合いが人種間で異なる可能性がある．

したがって，人種による肌の色や彫りの深さ等の顔特徴に影響を与える可能性のある環境要因として，明度，彩度，コントラスト，解像度を変動させて調査する．

明度は各ピクセルの 256 段階で表された輝度値 x_b を変動させる．明度の変化の度合い n_b を [-11,10] の範囲の整数とすれば，変化後の各ピクセルの輝度値を $\hat{x}_b$ として次式のように変化させる．

$$\hat{x}_b = \begin{cases} \min(0, \lfloor x_b - 25.6 * n_b \rfloor) & (n_b >= 0) \\ \max(255, \lfloor x_b - 25.6 * n_b \rfloor) & (n_b < 0) \end{cases} \quad (8)$$

彩度は，RGB 色空間から円柱モデルの HSV 色空間に変換し，256 段階で表された Saturation（彩度）値 x_s を変動させる．彩度の変化の度合い n_s を [-11,10] の範囲の整数とすれば，変化後の画像の彩度を $\hat{x}_s$ として次の式のよう変化させる．

$$\hat{x}_s = \begin{cases} \min(0, \lfloor x_s - 25.6 * n_s \rfloor) & (n_s >= 0) \\ \max(255, \lfloor x_s - 25.6 * n_s \rfloor) & (n_s < 0) \end{cases} \quad (9)$$

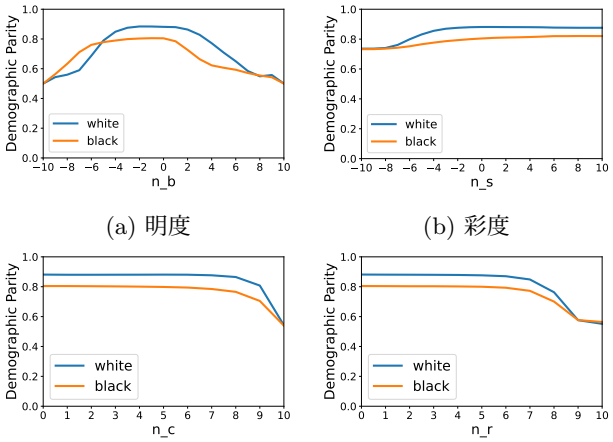
コントラストは，明度の最大と最小の差によって定義される．コントラストの変化の度合い n_c を [0,10] の範囲の整数とし，明度の最大値を b_{max}，最小値を b_{min} としたとき，n_c 段階目の変動におけるコントラスト c は $c = \min\{1, (b_{max} - b_{min}) * (10 - n_c)\}$ となる．各ピクセルの 256 段階で表された明度 x_c を変動させ，変化後各ピクセルの明度 $\hat{x}_c$ として次の式のよう変化させる．

$$\hat{x}_c = \lfloor \{(x_c - b_{min})/(b_{max} - b_{min})\}/c \rfloor + b_{min} \quad (10)$$

解像度は，画面解像度のことを指し，幅のピクセル数 × 高さのピクセル数 で定義される．解像度の変化の度合い n_r を [0,10] の範囲の整数とすれば，n_r 段階の変動における解像度は幅 $\min\{1, 10.8 * (10 - n_r)\}$ ピクセル，高さ $\min\{1, 12.4 * (10 - n_r)\}$ ピクセルとなる．

5 実験結果

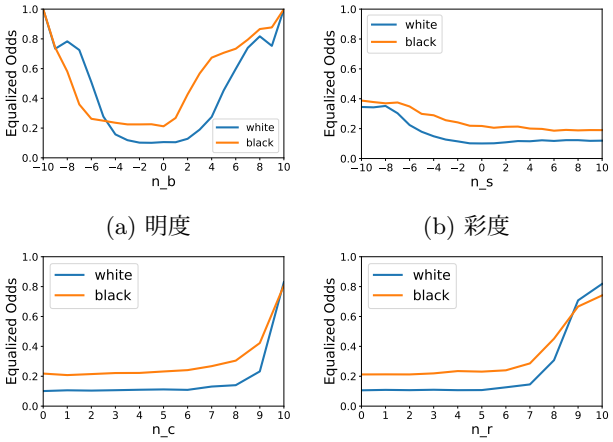
環境要因の変動に対する公平性の評価をデモグラフィックパリティ(図 3)，Δデモグラフィックパリティ(図 5)，ΔΔデモグラフィックパリティ(図 7)，等価オッズ (図 4)，



(a) 明度　(b) 彩度

(c) コントラスト　(d) 解像度

図 3: デモグラフィックパリティによる公平性評価



(a) 明度　(b) 彩度

(c) コントラスト　(d) 解像度

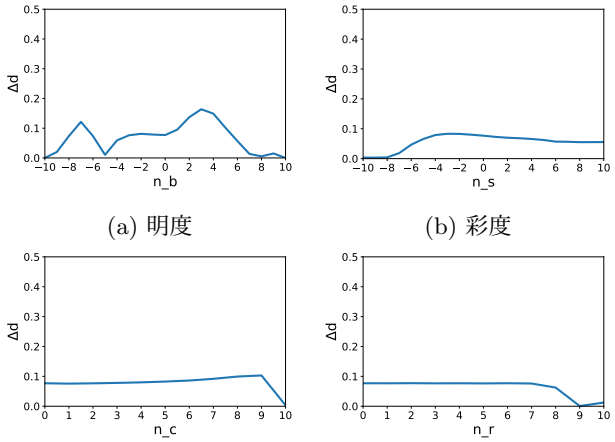
図 4: 等価オッズによる公平性評価

Δ等価オッズ(図 6)，ΔΔ等価オッズ(図 8) の 6 つの指標で検証した．縦軸はそれぞれの指標の評価値を示し，横軸は明度，彩度，コントラスト，解像度それぞれの環境要因の変動を明度，彩度は 21 段階，コントラスト，解像度は 11 段階で表している．

また，デモグラフィックパリティは式 (1) より認証精度に等しい．したがって，図 3 は環境要因の変動に対する各人種の認証精度の推移として評価することもできる．

図 3(b) より彩度を増大させた場合において，黒人は白人よりも緩やかに認証精度が低下した一方，白人は $n_s = -4$ において認証精度を大きく下げ，図 8(b) の $n_s = -8$ において ΔΔ等価オッズ は最も高い数値となっている．以上のことから，白人は黒人と比較して，色の変動に敏感であることが考えられる．

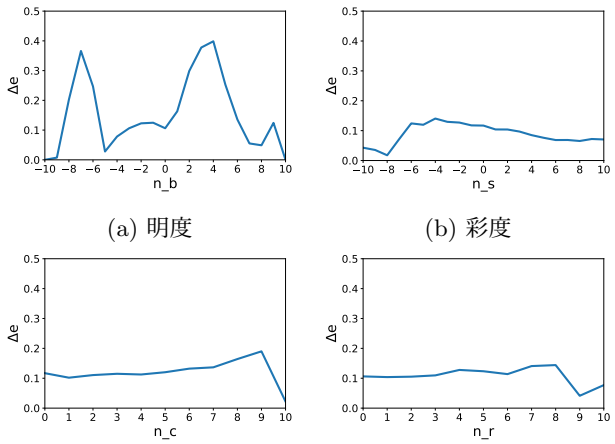
図 7，8 より明度が最も公平性に影響を与えることがわかる．さらに ΔΔ等価オッズ は ΔΔデモグラフィックパリティ よりも環境要因の変動に対して値が大きく変化することがわかる．



(a) 明度　(b) 彩度

(c) コントラスト　(d) 解像度

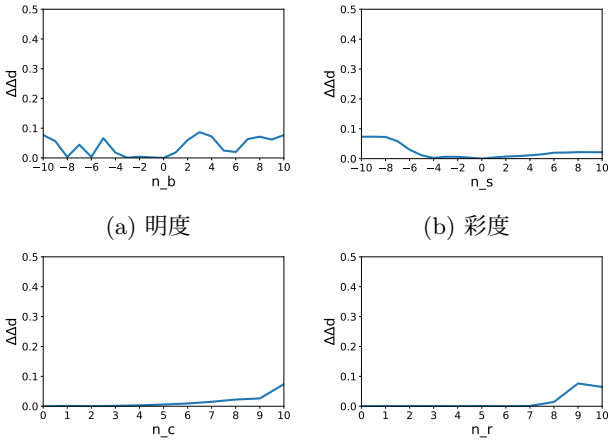
図 5: Δデモグラフィックパリティ による公平性評価



(a) 明度　(b) 彩度

(c) コントラスト　(d) 解像度

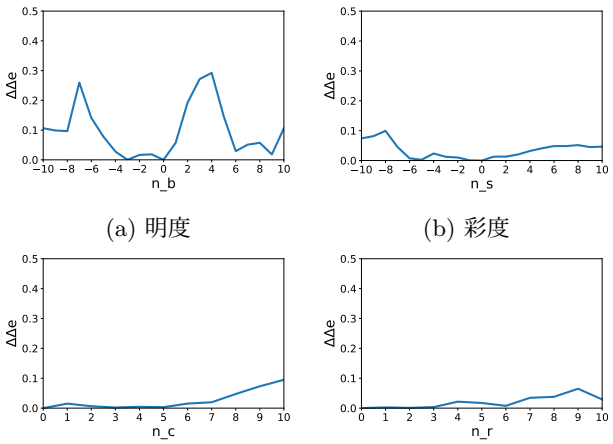
図 6: Δ等価オッズ による公平性評価



(a) 明度　(b) 彩度

(c) コントラスト　(d) 解像度

図 7: ΔΔデモグラフィックパリティ による公平性評価



(a) 明度　(b) 彩度

(c) コントラスト　(d) 解像度

図 8: ΔΔ等価オッズ による公平性評価

6　議論

6.1　デモグラフィックパリティと等価オッズについて

デモグラフィックパリティは人種間の認証精度としても評価することができるため理解しやすい指標であると考える．等価オッズは合致率に対する誤合致率の比を評価の値としているため，値が大きいほど認証精度が低いことを示す．しかし，図 3，4 に示す通り，デモグラフィックパリティと等価オッズのどちらにおいても，公平性を評価する場合は属性間の値の関係性を考慮する必要があるため，デモグラフィックパリティと等価オッズの指標だけでは公平性を評価することは難しいと考える．

6.2　Δ デモグラフィックパリティと Δ 等価オッズによる公平性評価について

デモグラフィックパリティや等価オッズと異なり，指標の値が 0 に近いほど公平であると評価することができる．この点からデモグラフィックパリティ，等価オッズの指標と比較し，公平性を評価することに適した指標であると言える．たとえば，図 5(b)，および図 6(b) より，$n_s = -8$ とすることで人種間の公平性が $n_s = 0$ の時の評価値と比べて改善されることがわかる．

6.3　ΔΔ デモグラフィックパリティと ΔΔ 等価オッズによる公平性評価について

ΔΔ デモグラフィックパリティおよび ΔΔ 等価オッズは，Δd_0 もしくは Δe_0 を基準とし，環境要因による変動量のみを評価することを目的とした指標である．Δ デモグラフィックパリティや Δ 等価オッズと異なり，評価の値が大きいほど環境要因が属性間の公平性に影響を与えていると評価することができる．一方，本指標の性質上，属性間の公平性を直接的に評価することはできない点に注意が必要である．したがって，ある変動地点での属性間の公平性を直接的に比較したい場合には Δ デモグラフィックパリティあるいは Δ 等価オッズを用いることが適切となる．たとえば $n_b = -5$ に着目すると，図 8(a)

より，ΔΔ 等価オッズは $n_b = 0$ と比較して高い値となっており，環境要因の変動が公平性に影響を与えていると評価することができる．一方，図 6(a) より，$n_b = -5$ における Δ 等価オッズは $n_b = 0$ と比較して低い値となっており，$n_b = 0$ と比較して人種間が公平であると評価される．

また，図 7，8 より ΔΔ デモグラフィックパリティと ΔΔ 等価オッズの公平性指標を比較すると，ΔΔ 等価オッズの方が環境要因の変動に対する評価値の変化量が大きい．これは ΔΔ 等価オッズは誤合致率と合致率の比から算出されるため，ΔΔ デモグラフィックパリティのように人種間の認証精度の差分を用いている場合と比較して変化量が大きくなりやすい点に起因すると考えられる．

6.4 環境要因が与える人種バイアスへの影響について

図 3(a) より明度の変動は人種バイアスに最も大きな影響を与えた．明度を増大させた場合の認証精度は $n_b = -5$ において黒人の認証精度を下回り，明度を減少させた場合も公平性指標に大きな変動が見られた．この原因については，肌の色による特性で，人種によって黒人であれば黒つぶれ，白人であれば白飛びの発生確率に差があるためと考えられる．図 8(a) より，明度を増大させた場合は $n_b = -7$ において ΔΔ等価オッズ が最も高く，明度を減少させた場合は $n_b = 4$ において ΔΔ等価オッズ が最も高いことがわかる．このことから，黒人は白人と比較して明度の変動による影響を受けやすいと考えられる．

また，本実験では，人種による顔の彫りの深さや骨格の違い等は，陰影などの顔特徴に起因することから，コントラストや解像度などの環境要因の変動が人種間の公平性に影響を与えることを想定していた．しかし，コントラストについては図 7(c) や図 8(c) からわかるように，ΔΔ デモグラフィックパリティや ΔΔ 等価オッズの変動が極めて少ないという結果が得られた．このことから，これらは公平性変動に影響を与えにくい環境要因であると言える．なお，$n_c > 7$ 以降で ΔΔ デモグラフィックパリティ，ΔΔ 等価オッズの変動が若干見られるが，これは変動が極めて大きくなった際に，人種を問わず識別精度が著しく低下することに起因する．

同様に解像度についても図 7(d)，図 8(d) より評価値は低い値となっており，人種間の公平性に与える影響は小さいことがわかった．

以上の公平性指標による実験結果から，実利用時の環境要因が人種バイアスを引き起こす可能性が示されたと考える．このような潜在的人種バイアスは，学習データの人種割合を統一するだけでは解消できない問題であり，人種に対して公平な顔認証システムか検証するためのデータセット等を用いて検証しただけでは公平かどうか真に判断することができないことを示している．

6.5 制限事項

本研究において使用したモデルは学習済みモデルの VGGFace [18] である．このため，VGGFace モデルの学習時点で学習データに偏りが生じており，これが環境要因変動への頑健性にも影響を与えた可能性がある．

このような影響を除くためには，学習データに含まれる属性の偏りを均一とした上で学習させたモデルを用いた評価結果を本稿の結果と比較分析する必要がある．

7 おわりに

本稿では，顔認証システムの公平性について，環境要因の変動を考慮した公平性評価を目的とした検討を行った．機械学習モデルの実利用時の環境要因が変動する際の公平性を評価する指標として，ΔΔ デモグラフィックパリティと ΔΔ 等価オッズを提案し，明度，彩度，コントラスト，解像度等の環境要因について，テストデータに変動を与え，提案した公平性指標により人種間の公平性に与える影響度を調査した．その結果，特に明度の変動において，変動の大きさにより人種間の認証精度の偏りが大きくなることがあることを示した．これは，人種バイアスに影響を与える潜在的要因が存在し，既存の公平性を評価するテストセットによる検証だけでは実利用時の潜在的人種バイアスが評価できないことを示している．また，データセットの人種割合の偏りを除くだけでは実環境における認証精度の偏りは解消されない可能性があることを示唆している．これらの結果をふまえ環境要因の変動を考慮することは，実利用時においても人種間の公平性が保たれた顔認証システムの実現へと繋がる．

本稿では学習済みモデルの VGGFace を使用したが，今後は，人種間のスコア正規化アプローチ等を行った人種バイアスを考慮したモデルを使用した場合においても実利用時の環境要因の変動によって潜在的人種バイアスが存在するかどうかの調査や，環境要因以外の顔画像データが持つ要因に起因する潜在的人種バイアスの調査も行っていく必要があると考える．さらに，環境要因を変動させたデータを学習時に水増しさせるなど，潜在的人種バイアスの軽減を考慮させるための手法について可能であるか調査も行っていく．

参考文献

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias: There ' s software used across the country to predict future criminals. And it ' s biased against blacks. 2016.

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on*

fairness, accountability and transparency, pp. 77–91. PMLR, 2018.

[3] 大阪大学. 第 3 部 海外の法規制及び社会動向, pp. 51–101. 国立国会図書館, 2019.

[4] Equal Credit Opportunity Act. `https://www.ftc.gov/enforcement/statutes/equal-credit-opportunity-act`, 1974. Accessed:2021/1/3.

[5] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

[7] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pp. 87–102. Springer, 2016.

[8] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 692–702, 2019.

[9] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Post-Comparison Mitigation of Demographic Bias in Face Recognition Using Fair Score Normalization. *Pattern Recognit. Lett.*, Vol. 140, pp. 332–338, 2020.

[10] Mei Wang and Weihong Deng. Mitigating Bias in Face Recognition Using Skewness-Aware Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9322–9331, June 2020.

[11] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Deep Imbalanced Learning for Face Recognition and Attribute Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, pp. 2781–2794, 2020.

[12] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, Vol. 21, No. 2, pp. 277–292, 2010.

[13] NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software. `https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software`, 2019. Accessed:2021/1/3.

[14] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. UNREGULATED POLICE FACE RECOGNITION IN AMERICA. `https://www.perpetuallineup.org/findings/racial-bias`, 2016. Accessed:2021/1/3.

[15] F. Kamiran and T.G.K. Calders. Classification with no discrimination by preferential sampling. In *Informal proceedings of the 19th Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn'10, Leuven, Belgium, May 27-28, 2010)*, pp. 1–6, 2010.

[16] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. Imperfect ImaGANation: Implications of gans Exacerbating Biases on Facial Data Augmentation and Snapchat Selfie Lenses. *arXiv preprint arXiv:2001.09528*, 2020.

[17] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, p. 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[18] Refik Can Malli. keras-vggface. `https://github.com/rcmalli/keras-vggface`, 2021. Accessed:2021/1/3.

[19] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–10, 2020.