顔認証における公平性評価の一検討

大木　哲史[1,2,a)]　荒井　ひろみ[2]

概要：顔認証技術は，特に欧米において犯罪捜査を目的とした利用が進んでいるが，一方で近年，学習データの偏りによって，有色人種や性別などのセンシティブ情報に依存して顔認証技術の精度が大きく変動することが社会問題となりつつある．本研究では，顔認証における公平性について取り上げる．例えば犯罪者リストとの照合など人生に重大な影響を与える用途において．ある特定のセンシティブ情報をもつグループが犯罪者に誤認識されやすいような状況は不公平であると言える．本研究ではこのような不公平さについて，特に表現学習に基づく汎用的な顔認証システムを取り上げ，その公平性評価方法について検討，評価を行った．MORPH データセットを用いた実験により，デモグラフィックに偏りのある学習セットにより作成された顔特徴抽出器を用いた場合に，認証精度の偏りが生じることや，偏りのあるテストセットを用いることで意図的に高い公平性評価値を算出可能である，属性ごとのしきい値を設定することで認証精度への影響を軽減できることなどが示された．

キーワード：顔認証，公平性，生体認証

A Study on Fairness Evaluation for Face Recognition Systems

Tetsushi Ohki[1,2,a)]　Hiromi Arai[2]

Abstract: Face recognition technology is increasingly being used for criminal investigations, especially in Europe and the United States. However, it has become a social problem that the accuracy of face recognition technology varies greatly depending on sensitive information such as race and gender due to training data bias. For example, in a life-critical application such as matching with a criminal list, it is unfair that criminals easily misrecognize a group with certain sensitive information. This paper discusses and evaluates the fairness evaluation method of a general-purpose face recognition system based on representation learning. The MORPH dataset experiments showed that authentication accuracy bias occurs when using a face feature extractor created with a demographically biased training set. Furthermore, we showed that it is possible to intentionally calculate a high fairness evaluation value using a biased test set. The effect on authentication accuracy can be reduced by setting a threshold value for each attribute.

Keywords: Face Recognition, Fairness, Biometrics

1. はじめに

個人を特定できる身体的または行動的な特徴を用いて認証を行う生体認証技術の利用が進んでいる．特に顔認証は非接触な認証が可能であること，カメラなどの安価なセンサーのみで実現できるなどの利点があり様々な場面で応用されている [4]．しかし顔認証には個人の情報を用いるがゆえの懸念も存在する．例えばプライバシーの懸念から顔データの収集や公共の場所における顔認証の利用が制限される傾向にある [13]．加えて近年顔認証における公平性の懸念が提起されている．

近年人工知能の公平性については，アルゴリズムによる自動判定が人種や性別などの情報に関して差別的な振る舞いをすることが問題視されている．人工知能の公平性は，

[1] 静岡大学
Shizuoka University
[2] 理化学研究所 革新知能統合研究センター
RIKEN AIP
[a)] ohki@inf.shizuoka.ac.jp

社会的公平性の観点からその関与を排除すべき属性をセンシティブ特徴とし，センシティブ特徴にまつわるバイアスについて様々な側面から議論されている．

顔認証における公平性に関する事例としては．学習データの偏りによって，有色人種や性別などに起因して顔認識の精度が大きく変動することが社会問題となったことが挙げられる．この問題は，米国で 2020 年 5 月に発生した白人警察官による黒人男性殺害事件に端を発する人種差別抗議を受け，IBM，Amazon，Microsoft といった大手テック企業が相次いで同社の顔認証技術を出荷停止する等の事態へと発展している [13].

これまで顔認証の公平性については性別や人種といったデモグラフィック属性において，多数属性 (その属性値を持つ人がデータセット中の多数を占める) 集団と少数属性 (その属性値を持つ人がデータセット中で少数である) 集団間でクラス誤分類確率に差が発生する点について主に議論がなされてきた．特に，生体認証において犯罪捜査は主要なアプリケーションの 1 つであるが，そのような分野においては，あらかじめ登録された犯罪者データベース（ウォッチリスト）に含まれる人物であるかを確認する作業が発生する．ここで，ウォッチリストが不均衡な場合，マイノリティグループの識別率が大幅に増加することなどが報告されている [3]．これはいわゆる 1 対多の identification シナリオ において，たとえば黒人をマイノリティグループとした場合に，黒人同士の類似度が他のグループと比較して著しく高く出力されることで，自分以外の人物が候補として出力されるシナリオを想定していると考えられる．

ここで，文献 [3] をはじめとする，顔認証におけるこれまでの検討が対象としてきたウォッチリスト照合では，ウォッチリストに登録された人物であるかを判定する多クラス識別器が偏った学習データにより学習されている場合の影響について検討を行っている．一方で，汎用的な顔認証においては，登録するユーザの数は限定せずに，学習データを用いて表現学習を実施することで得た顔特徴の分散表現を顔特徴量として用いることが多い．このような表現学習に用いる学習データに偏りが生じた場合に，照合・識別結果に与える影響がどのようなものになるのかは未だに詳細な検討が成されていない．そこで，本研究では表現学習に基づく汎用的な顔認証システムにおける公平性について取り組み，その評価方法について検討，評価を行う．

本研究の貢献は以下である．

- 顔照合・識別を含む顔認証システムのシナリオを整理するとともに，各シナリオにおける顔認証システムの公平性評価方法を提案した．
- MORPH データセットを用いて $\Delta\mathrm{FMR}(\tau)$, $\mathrm{FPIR}(\tau, N, s)$ といった指標を評価・比較することで，学習データの偏りが顔特徴抽出器および顔照合・識別精度に与える影響を示した．

2. 関連研究

人工知能の公平性について，これまで機械学習・データマイニングを中心に様々な角度から論じられてきた．機械学習においては，人種や性別など社会的公平性の観点からその関与を排除すべき属性をセンシティブ特徴とし，様々な公平性の定義が提案されてきた．

顔認証については，まず顔画像データを用いて顔の持ち主の特徴を分類する顔識別タスクについての不公平性が指摘されてきた．ベンチマークデータのサンプル数における性別および肌色の偏りおよび商用の顔識別システムにおける性別分類などのタスクの予測精度の偏りが報告されている [1], [9]．また既存の公開顔データにおける顔画像データの偏りについても指摘されており，人種ごとのサンプル数のバランスをとったデータセットを作成し，それをもとに学習すると人種，性別，年齢の分類タスクで高い精度を達成すると報告されている [8]．また顔画像における性別分類タスクにおける学習データの偏りの影響の分析し，不公平なモデルを検出する手法が提案されている [10]．これらの研究は公平性について重要な議論であるが，多くの場合は属性推定を目的としており，我々の扱う個人の照合・識別とは異なるタスクであるため，評価指標や結果をそのまま援用することができない．

顔識別システムを対象とした公平性に関する検討としては，識別システムを対象として，センシティブ特徴で条件付けられた FMR についての公平性についてウォッチリストにおけるマイノリティグループの識別率が大幅に増加するという報告がなされている [3]．ウォッチリストは犯罪捜査の目的下では現実的なシナリオではあるが，より汎用の顔照合・識別を想定した場合には，異なる観点の評価シナリオが必要となる可能性がある．また，照合システムに関しては，顔認識における偏りの影響を軽減するための教師なし fairscore 正規化アプローチが提案されている [11], 本手法は照合時に入力される画像の属性を判定することで，属性に有効なしきい値を用いた照合処理を可能とする手法であるが，識別シナリオにおいては多数のトランザクションが発行されることから，属性判定等の処理が困難となる可能性がある．

3. 顔認証システムの評価

3.1 生体認証システムの認証シナリオ

生体認証において，利用者による肯定的な身元確認要求に対し，提示された生体情報から生成した特徴を用いて登録テンプレートとの比較を行い，身元確認要求に対する成否を返答する処理を照合（Verification）と呼ぶ．これに対し，登録データベースの検索を実行し，0 個または 1 個以上の身元情報からなる候補リストを返す場合にはこれを識別（Identification）と呼ぶ．なお，識別処理には，すべての

潜在的利用者がシステムに登録されている登録者限定識別（Closed-set Identification）および，すべての潜在的利用者がシステムに登録されているとは限らない識別（Open-set Identification）が存在する．ここで照合・識別両者を含むシステムを単に認証システムと呼ぶこととする．

3.2 深層学習を用いた顔認証システム

顔認証システムの深層学習による構成について考えるために，まず顔画像 $\boldsymbol{x}$ に対応する人物ラベル $y \in [1, n]$ を出力とする n クラス識別のニューラルネットワーク $F(\boldsymbol{x}) = y$ を考える．ここで，F を classifier 関数（softmax など）を含む完全なニューラルネットワークとして定義する．さらに，$Z(\boldsymbol{x})$ を classifier 関数を除いたネットワークの出力として定義すればこの n クラス識別ニューラルネットワークは式 (1) のように定義できる．

$$F(\boldsymbol{x}) = \arg\max(\texttt{classifier}(Z(\boldsymbol{x})) = y. \qquad (1)$$

顔認証システムは，認証システムに登録するユーザの増減が頻繁に発生すること，特定ユーザの画像のみを大量に収集することが困難であること，といった理由から，$F(\boldsymbol{x})$ を用いた多クラス分類器を認証システムとして使用することは難しい．このため事前に一定数の学習データから多クラス分類器 $F(\boldsymbol{x})$ を学習し，その後，$Z(\boldsymbol{x})$ を顔特徴抽出器として用い，その出力として得られるベクトル間の類似度により判定を行うことが一般的である．なお，$F(\boldsymbol{x})$ の学習において，一般的な物体認識では同一クラス内における $Z(\boldsymbol{x})$ 間の二乗誤差が最小化されるように $F(\boldsymbol{x})$ 内の classifier 関数が定義されることが一般的であるが，顔認証においては，ArcMaginLoss[2] など，コサイン類似度を最小化するような classifier 関数が用いられることが多い．

3.3 顔認証システムの評価尺度

ISO/IEC 19795-1[6] において，生体認証の照合性能を評価するための評価尺度は，誤非合致率（false non match rate, FNMR）や誤受理率（false match rate, FMR）を用い，識別性能の評価尺度は，誤拒否識別率（false-negative identification rate, FNIR）および誤受入識別率（false-positive identification-error rate, FPIR) を用いることが推奨されている．

ここで，顔画像のペア $(\boldsymbol{x}_r, \boldsymbol{x}_q)$ について考える．顔認証システムによって特徴ベクトルを抽出した後，類似度関数 $\mathcal{L}$（たとえば，コサイン類似度）により，2 つの顔画像間の類似度を算出することが可能である．ここで $\boldsymbol{x_r}$ および $\boldsymbol{x_q}$ をそれぞれユーザ r および q の生体情報とし，類似度関数 $\mathcal{L}$ の出力がしきい値 τ 以上となった場合にユーザ r および q が同一人物と判定されるとすれば，FNMR および FMR は次式で定義できる．

$$\mathrm{FNMR}(\tau) = \Pr[r = q | \mathcal{L}(x_r, x_q) < \tau] \qquad (2)$$

$$\mathrm{FMR}(\tau) = \Pr[r \neq q | \mathcal{L}(x_r, x_q) \geq \tau] \qquad (3)$$

また，FNMR と FMR が等しくなる時の誤り率を EER と呼び，次式で定義する．

$$\mathrm{EER} = \mathrm{FNMR}(\tau_*) = \mathrm{FMR}(\tau_*) \qquad (4)$$

ここで，τ_* は EER を達成するしきい値 τ であり，次式で定義される．

$$\tau_* = \arg\min_{\tau}(|\ \mathrm{FNMR}(\tau_*) - \mathrm{FMR}(\tau_*)\ |) \qquad (5)$$

なお，τ が離散的に与えられた場合，厳密に FMR と FNMR が一致する点を決定することが困難な場合がある．その場合は近似として，以下の式で計算される HTER（Half Total Error Rate）が用いられる．

$$\mathrm{HTER}(\tau) = \frac{1}{2}(\mathrm{FMR}(\tau) + \mathrm{FNMR}(\tau)) \qquad (6)$$

識別システムの性能である FNIR, FPIR のうち，特に FPIR は認証システムへの登録ユーザ数 N の影響を受ける．ここで，識別システムは，しきい値 τ を超えた登録ユーザのリストを返すシステムとして定義される．この時，FNIR および FPIR は返却されるリストに該当ユーザが含まれる確率となるので，FNIR および FPIR は N を用いて次の式で示される [5].

$$\mathrm{FNIR}(\tau, N) = \mathrm{FNMR}(\tau, 1) \qquad (7)$$

$$\mathrm{FPIR}(\tau, N) = 1 - (1 - \mathrm{FMR}(\tau))^N \qquad (8)$$

ISO/IEC 19795-1:2021[6] では，FPIR および FNIR を FTA（Failure to Acquire Rate）含む指標として定義している．FTA はセンサ入力画像の品質等を理由として認証プロセスそのものが実行されない確率であるが，本稿においては全ての画像に対し認証プロセスを実施すると仮定し，単に FTA=0 とした上記の定義を採用する．

異なる照合システム間の性能比較にあたっては，EER を用いる場合と，EER 以外を評価尺度として用いる場合がある．EER 以外を評価尺度としたい場合には，多くの場合は利便性要件である FNMR を複数システム間で同一とした上で，FMR の性能を比較することが一般的である．同様に，識別システムにおいても，FNIR を固定した上で，FPIR の比較を行うが，識別システムにおいては登録者数 N の増加に伴い FPIR も急速に増大する傾向があるため，N を考慮した FPIR の評価を行う必要がある．

3.4 顔認証システムの公平性評価

顔認証システムの公平性評価について，ここでは Verification（照合）シナリオと Identification（識別）シナリオの 2 つの場合に分けて考える．Verification シナリオにおいては，センシティブ属性 s の有無によって生じる，FMR(τ) の

差を評価し，Identification シナリオにおいては FPIR(τ,N) 生じる差を評価する．ここで，単純のために，センシティブ属性を持つグループ s_1 と持たないグループ s_2 の 2 つに全ユーザが分割できると仮定すれば，照合時に評価する FMR の差 $\Delta\,\mathrm{FMR}(\tau)$ は，

$$\Delta\,\mathrm{FMR}(\tau) = \mathrm{FMR}(\tau, s_0) - \mathrm{FMR}(\tau, s_1) \tag{9}$$

ここで，$\mathrm{FMR}(\tau, s)$ は，

$$\mathrm{FMR}(\tau, s) = \Pr[r \neq q | s(\boldsymbol{x_r}, \boldsymbol{x_q} \geq \tau), \forall_r \in \mathcal{U}_s, \forall_q \in \mathcal{U}_s] \tag{10}$$

ただし，$\mathcal{U}_s$ はセンシティブ属性 s を持つグループに属するユーザの集合である．また，識別時に評価する FPIR の差 $\Delta\,\mathrm{FPIR}(\tau, N)$ は，

$$\Delta\,\mathrm{FPIR}(\tau, N) = \mathrm{FPIR}(\tau, N_{s_0}, s_0) - \mathrm{FPIR}(\tau, N_{s_1}, s_1) \tag{11}$$

として求められる．ここで N_{s_0} および N_{s_1} は登録ユーザ N 人中にセンシティブ情報 s_0 もしくは s_1 を持つユーザの数であり，$\mathrm{FPIR}(\tau, N, s)$ は次式で与えられる，

$$\mathrm{FPIR}(\tau, N, s) = 1 - (1 - \mathrm{FMR}(\tau, s))^N \tag{12}$$

4. 評価実験

本稿で 3 章で提案した評価指標を用いた公平性評価を MORPH データセットを対象とした顔照合・顔識別実験を通じて実施する．センシティブ属性ごとに偏りのあるデータセットを作成し，これにより顔認証システムの学習・テストを行い，照合・識別精度への影響を 3 章で検討した評価指標によって確認する．ここで，データセットはセンシティブ属性について，多数属性，少数属性の 2 つのグループに分類されるものとする．

4.1 実験条件

本実験では，実験用データセットとして UNCW が提供する MORPH データセット [12] を用いた．MORPH は加齢に伴う顔画像の変化を評価することを目的としたデータセットであるが，顔画像とともに性別，年齢，人種の属性情報が記録されていることから，これらによる影響を評価するデータセットとしても広く利用されている．

MORPH データセットは，13,617 名の 55,134 枚の顔画像から構成される．ここで，個人内の撮影枚数が 1 枚のみの人物 457 名については，本人間照合の実施が困難であるため，本実験の対象データから除外した．また，特に人種属性に関して，同一人物に異なる属性が誤って設定されている人物が存在した．このような人物 34 名を本実験の対象データから除外した．これらの処理を行った後，合計 54,531 枚の画像を実験に使用した．表 1 に使用した画像データセットの属性ごとの画像数の分布を示す．データセット内は，アフリカアメリカン（B）男性（M）の画像が大きな割合で含まれていることがわかる．

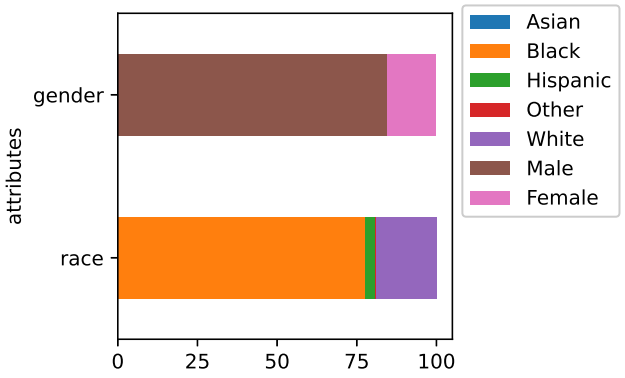


図 1: MORPH データセット内の属性分布

また，各画像から背景や服装等の影響を取り除くために，顔周辺領域の切り出しを MTCNN を行った．これらの画像から ResNet18+ArcFace [2] を用い顔認証モデルを学習した（以下，単に ArcFace と呼ぶ）．テスト時は，ArcFace の手法に従い，2 枚の画像それぞれを学習したモデルに入力することで，埋め込み層の出力から特徴ベクトルを得た後，この特徴ベクトル間のコサイン類似度を用いて認証を行った．

4.1.1 偏りのあるモデルの作成

本実験では，データセットの 90 %をモデルの学習に，10 %をテストに使用した．学習したモデルは多クラス識別器ではなく，表現学習に基づく顔特徴抽出器として用いられる．このため学習セットとテストセットの間には人物の重複は設けなかった．モデルの作成においては，学習データにおけるセンシティブ情報構成に偏りがある場合と均等な場合を想定し，MORPH に含まれる属性のうち，性別（男性 [M]，女性 [F]）および人種（アフリカアメリカン [B]，ヨーロピアン [W]）について，学習データに含む属性の構成割合に (1) 著しい偏りがある場合（90:10），(2) 偏りがある場合（75:25），(3) 均等な場合（50:50）として学習を行った 3 つのモデルを作成した．以降の実験では，それぞれ，性別に関して偏りを生じさせたモデルを gender_{90-10, 75-25, 50-50}, 人種に関して偏りを生じさせたモデルをそれぞれ race_{90-10,75-25,50-50} のように記載する．なお，図 1 から分かるように，性別に関しては男性 [M] が，人種に関してはアフリカアメリカン [B] が多数属性となっており，本稿でも各モデルを構成する多数属性を常に性別に関しては [M]，人種に関しては [B] として作成した．

4.1.2 評価シナリオ

Identification シナリオにおいては，テストセットのうち一部の画像を登録データとした上で，残りのデータを用いて識別実験を行う必要がある．本実験においては，テスト

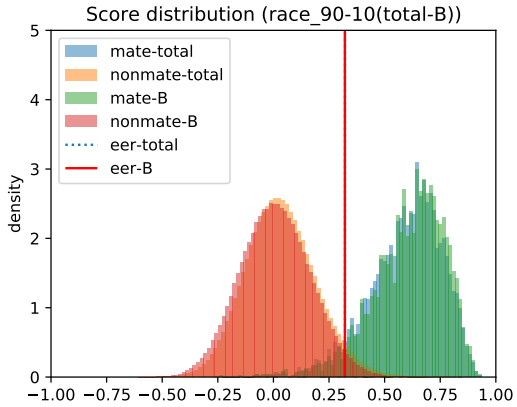


図 2: 類似度分布の比較
(モデル race_90-10 / total-B)

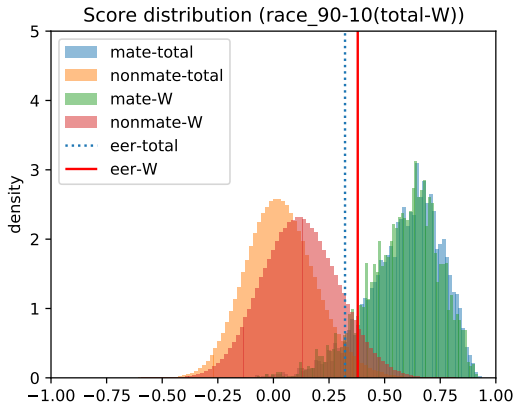


図 3: 類似度分布の比較
(モデル race_90-10 / total-W)

データのうち N 人から 1 枚ずつランダムに画像を選択する．選択した N 枚の画像から抽出した顔特徴量を参照用データし，選択した N 人を除いたテストデータ中の残りの人物の画像を用いて識別実験を行った．ただし，登録データに含まれる各属性は均等となるようにした．つまり，式 (11) において $N_{s_0} = N_{s_1}$ とした．なお，登録データとして選択する画像により識別精度が変動する可能性を考慮して，テストセットから N_{s_0} および N_{s_1} の選択パターンに関して 10-fold cross validation を実施した上で，識別率 FPIR(τ, N, s) の平均と分散を算出した．

4.1.3 テストセットの選択

照合・識別いずれのテスト時においても，テストに用いるデータをランダムに選択した場合，学習データと同様に属性に偏りが生じ，これによる照合・識別精度への影響が発生する可能性がある．このため，本実験では，テスト時に各属性から均等な割合でテストデータを選択する balanced 条件，モデル作成時と同じ比率でテストデータを選択する imbalanced 条件の 2 つの条件下で検証を行った．

4.1.4 適応的しきい値判定法

図 2 は全ユーザ間で算出したスコア分布とアフリカアメリカンユーザ間のみで算出した類似度分布である．race_90-10 モデルにおいては，学習データの 9 割がアフリカアメリカンから構成されているため，類似度分布の差はほぼないことがわかる．一方，図 3 は全ユーザ間で算出した類似度分布のヨーロピアンユーザ間のみで算出した類似度分布である．こちらに関しては，特に nonmate（他人間）スコアにおいて，大きな違いがあることがわかる．ここで，各図において，赤い実線を属性内照合スコアから算出した EER を達成する際のしきい値（eer-W/B），青い破線を全ユーザ間照合スコアから算出した EER を達成する際のしきい値（eer-total）として示している．特に図 3 においては，しきい値 eer-total としきい値 eer-W の間に大きな差があること，また，属性 [W] 間の他人間類似度分布（nonmate-W）が eer-total よりも大きく右に分布していることがわかる．このことから，属性 [W] は，race_90-10 モデルを特徴抽出器として利用する顔認証モデルに対して著しく他人受入（False Match）を発生しやすいユーザー群であることがわかる．

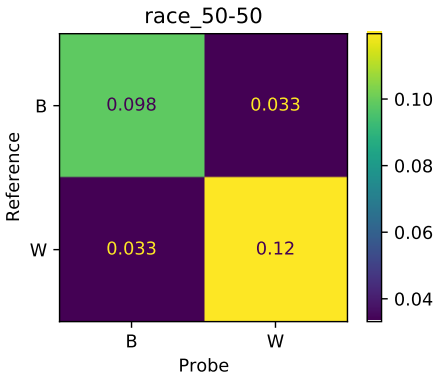


図 4: 属性間の FMR(モデル race_50-50)

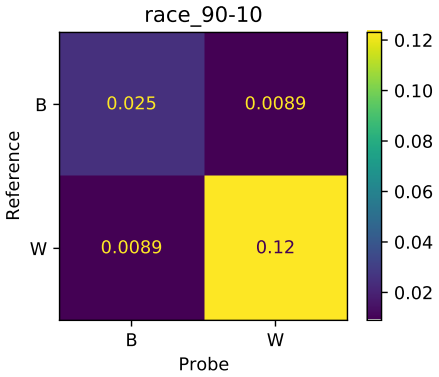
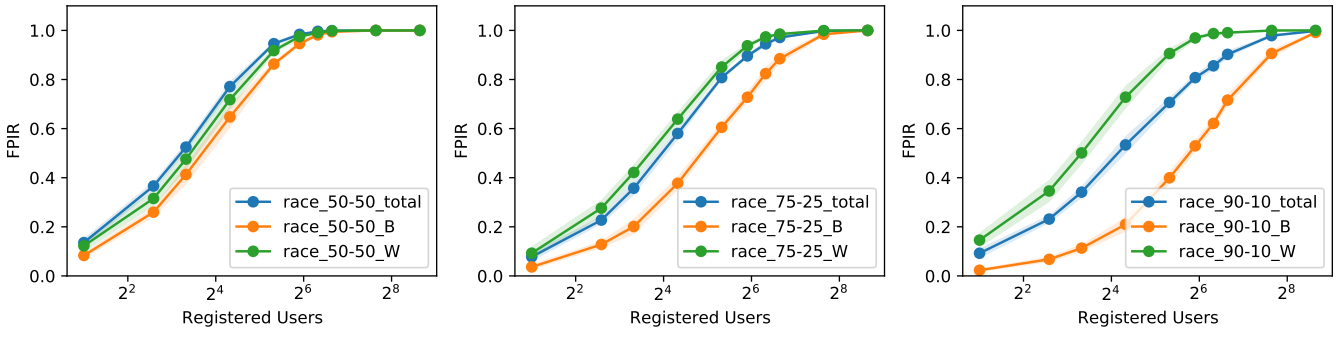
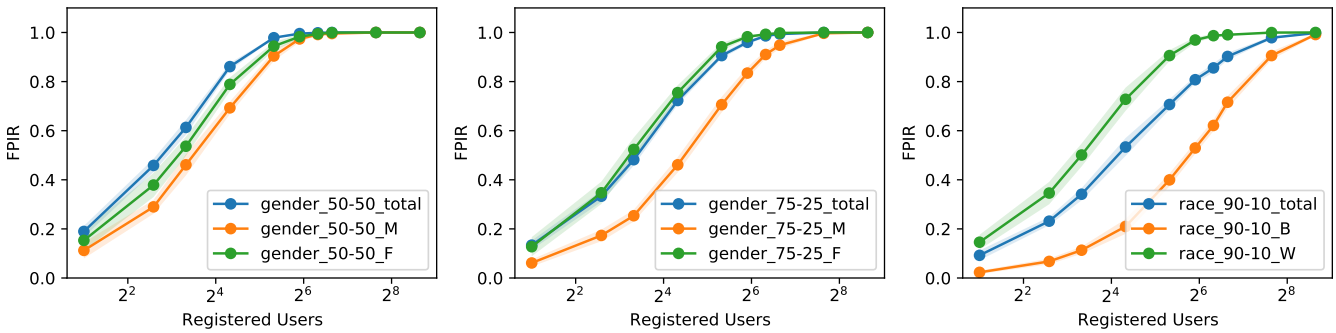


図 5: 属性間の FMR(モデル race_90-10)

このような他人受入の発生は，FPIR の算出にあたって，図 3 中の青い波線，すなわち照合実験により求めた τ_* を属性に関わらず全ユーザ共通のしきい値として使用することに起因する．したがって，事前に得られる学習データを用いて，次のような判定方法が可能であると考えられる．
(1) 事前学習データから属性ごとに最適な τ_* を算出する
(2) 判定時には登録データの属性に応じて事前に設定した



(a) race_50-50 **(b)** race_75-25 **(c)** race_90-10



(d) gender_90-10 **(e)** gender_75-25 **(f)** gender_90-10

図 6: FPIR(τ, N, s) の評価結果

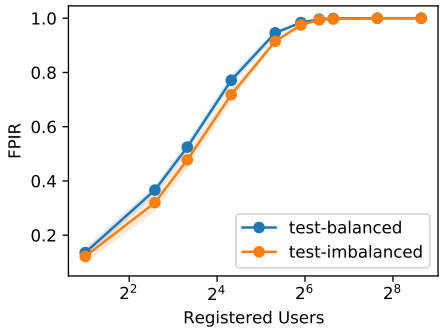


図 7: テストセットの違いによる FPIR の偏り（race_50-50 モデル）

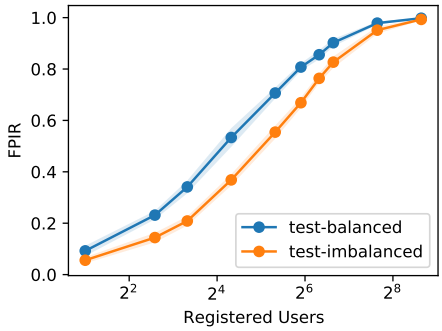


図 8: テストセットの違いによる FPIR の偏り（race_90-10 モデル）

しきい値 τ_* を使用する

このような手法は適応的しきい値判定法と呼ばれ，生体認証においては個人間における認証精度の偏りを軽減する目的で用いられてきた [7]．本実験では，各属性間の照合実験により，属性ごとの EER およびしきい値 τ_* を算出し，これを各属性用しきい値として用いることで，属性間 FPIR の偏りが改善されるかどうかについて検証する．

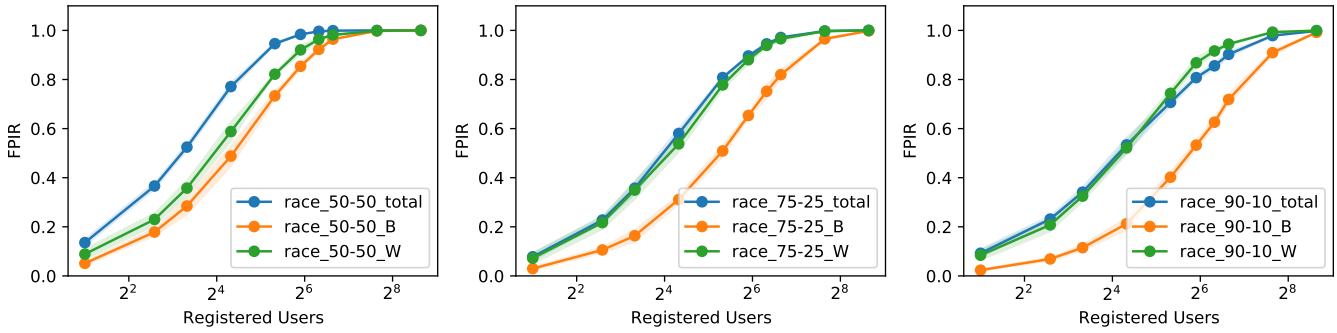
4.2 実験結果

4.2.1 照合実験

まず，図 4 および図 5 にセンシティブ属性を人種とし，学習データのデモグラフックに偏りがない場合（モデル race_50-50 を用いた場合）と偏りがある場合（モデル race_90-10），それぞれの照合シナリオにおける誤り率 FMR(τ, s_1) および FMR(τ, s_0) を示す．ここで，本実験において多数属性は [B] としていることから，$s_1 = B, s_0 = W$ である．また，τ の値は，属性 [B] および [W] を含めた全てのユーザを用いた照合実験を行い，式 (5) により決定した τ_* を用いた．

たとえば，図 5 は，90%が属性 [B]，10%が属性 [W] から成る学習セットで顔特徴抽出器を作成し，その後，テストセットに含まれるデータ間で他人間照合を行った場合の誤り率を示したものと言える．1 行 1 列目の値は，テストセットに含まれる属性 [B] 同士の他人間照合の類似度がしきい値 τ_* を超える確率が 0.025 であることを示している．

両シナリオにおける ΔFMR(τ, s) を比較すると，図 4 のシナリオにおいては ΔFMR(τ)=0.12 - 0.098=0.022 であるのに対し，図 5 のシナリオにおいては，ΔFMR(τ)=0.12



(a) race_50-50(adapted) **(b)** race_75-25(adapted) **(c)** race_90-10(adapted)

図 9: 識別実験の結果（適応的しきい値設定）

- 0.025=0.095 と $\Delta\mathrm{FMR}(\tau)$ がデモグラフィックの偏りに応じて増大していることがわかる．

4.2.2 識別実験

図 6(a)∼(f) にそれぞれ，人種および性別を対象とした識別実験の結果として各条件における $\mathrm{FPIR}(\tau, N, s)$ を示す．ここで，各図において横軸は識別システムへの登録ユーザ数 N，縦軸は $\mathrm{FPIR}(\tau, N, s)$ である．図 6(a)∼(f) それぞれの評価におけるしきい値 τ は対応するモデルの照合実験時に算出した τ_* を用いた．4.1.2 項で述べたように，識別実験は N に関する 10-fold cross validation を実施しており，これにより得られた $\mathrm{FPIR}(\tau, N, s)$ の平均値および平均 $\pm 1\sigma$ の範囲を示してある．また，図中の凡例において，モデル名に続いて記載された B/W/total は $\mathrm{FPIR}(\tau, N, s)$ 算出の対象とした属性名である．たとえば，属性 [B] ($s = s_1$) であれば，属性 [B] を持つユーザ間の FPIR，すなわち $\mathrm{FPIR}(\tau, N, s_1)$ を示す．また，'total' に関しては属性を問わない全ユーザ間の FPIR であり，$\mathrm{FPIR}(\tau, N)$ を示す．

本実験では，$\Delta\mathrm{FPIR}(\tau, N)$ を示す代わりに，複数条件の $\mathrm{FPIR}(\tau, N, s)$ を N の増加に伴う特性として示すことで，視覚的に $\Delta\mathrm{FPIR}(\tau, N)$ を評価した．人種，属性いずれの識別実験結果からも，特徴抽出器作成における属性の偏りが，FPIR 性能の偏りに影響していることが見て取れる．たとえば，登録ユーザ数 $N = 20$ の点に着目すれば，モデル race_50-50 においては $\Delta\mathrm{FPIR}(\tau, 20) = \mathrm{FPIR}(\tau, 20, s_0) - \mathrm{FPIR}(\tau, 20, s_1)$=0.723 - 0.651=0.072 であるのに対し，モデル race_90-10 においては $\Delta\mathrm{FPIR}(\tau, 20)$=0.721 - 0.209=0.512 と大幅に増加していることがわかる．

4.2.3 テストデータの選択

図 7 および図 8 にテストデータを 4.1.3 節で述べた balanced 条件とした場合の $\mathrm{FPIR}(\tau, N)$，および imbalanced 条件とした場合の $\mathrm{FPIR}(\tau, N)$ を示す．なお，選択するセンシティブ属性としてここでは人種を例に挙げ，比較対象として，同一属性間だけでなく，属性 [B]-[W] といった他属性間の識別トランザクションを含む識別タスクでの性能を取り上げる．

特に学習データのデモグラフィックに偏りがある図 8 において，$\mathrm{FPIR}(\tau, N)$ が異なる結果となっていることがわかる．これは，照合実験の図 5 で示したように，属性間の誤り率には偏りが生じており，このため，相対的に誤り率の高い [W]-[W] 間の比較が，デモグラフィックに偏りが生じた結果少なくなったことに起因すると考えられる．また，これはテストセットを恣意的に，たとえば多数属性を多くテストセットに含めることで，意図的に公平性指標である $\mathrm{FPIR}(\tau, N)$ を高く見せることが可能であることを示している．

4.2.4 適応的しきい値判定法

最後に，4.1.4 項で述べた適応的しきい値判定法により属性ごとのしきい値を設定した場合の識別実験結果を図 9 に示す．しきい値として，全テストデータから算出した τ_* を用いた total の結果を total とし，属性 [B]，属性 [W] については，4.1.4 項の手法にしたがってしきい値を設定した．適応的しきい値を用いない場合の評価結果である図 6(a)-(c) と比較すると，特にデモグラフィックに偏りが発生している 75-25,90-10 のモデルにおいて，属性 [W] の FPIR が適応的しきい値判定法により改善し，total と比較しても低い FPIR を達成していることがわかる．

5. まとめ

本稿では，近年公平性の懸念が提起される顔認証について，特に表現学習に基づく汎用的な顔認証システムを取り上げ，その公平性評価方法について検討，議論を行った．MORPH データセットを用いた実験では，表現学習に用いられるデータセットの偏りもまた，ウォッチリスト識別等のこれまで行われてきた議論と同様に認証精度の偏りを生じ得ることが示された．さらに，学習セットだけでなく評価に用いるテストセットを恣意的に作成することで，意図的に高い公平性評価値を算出可能であることも同時に示された．また，比較するテンプレートの属性に応じてしきい値を設定することで，顔特徴抽出器に存在する偏りの影響

を軽減できることが示された．今後の課題として，しきい値の最適化について公平性の観点からさらに検討を進め，顔認証における公平性の評価および改善方法を明らかにしていくことが挙げられる．

参考文献

[1] Buolamwini, J. and Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification, *Conference on fairness, accountability and transparency*, PMLR, pp. 77–91 (2018).

[2] Deng, J., Guo, J., Xue, N. and Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699 (2019).

[3] Drozdowski, P., Rathgeb, C. and Busch, C.: Demographic Fairness in Face Identification: The Watchlist Imbalance Effect, *arXiv preprint arXiv:2106.08049* (2021).

[4] Garvie, C.: *The perpetual line-up: Unregulated police face recognition in America*, Georgetown Law, Center on Privacy & Technology (2016).

[5] Grother, P. and Phillips, P.: Models of large population recognition performance, *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. II–II (online), DOI: 10.1109/CVPR.2004.1315146 (2004).

[6] ISO/IEC JTC 1/SC 37 Biometrics: ISO/IEC 19795-1:2021 Information technology - Biometric performance testing and reporting - Part 1: Principles and framework, , available from ⟨https://www.iso.org/standard/73515.html⟩ (accessed 2021-08-22).

[7] Jain, A. and Ross, A.: Learning user-specific parameters in a multibiometric system, *Proceedings. International Conference on Image Processing*, Vol. 1, pp. I–I (online), DOI: 10.1109/ICIP.2002.1037958 (2002).

[8] Karkkainen, K. and Joo, J.: FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558 (2021).

[9] Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J. and Denton, E.: Saving face: Investigating the ethical concerns of facial recognition auditing, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 145–151 (2020).

[10] Serna, I., Peña, A., Morales, A. and Fierrez, J.: InsideBias: Measuring bias in deep networks and application to face gender biometrics, *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3720–3727 (2021).

[11] Terhörst, P., Kolf, J. N., Damer, N., Kirchbuchner, F. and Kuijper, A.: Post-comparison mitigation of demographic bias in face recognition using fair score normalization, *Pattern Recognition Letters*, Vol. 140, pp. 332–338 (online), DOI: 10.1016/j.patrec.2020.11.007 (2020).

[12] Wilmington, U.: MORPH Facial Recognition Database, , available from ⟨https://uncw.edu/oic/tech/morph.html⟩ (accessed 2021-08-23).

[13] 岸本充生：続・生体認証技術の動向と活用，，入手先 ⟨https://elsi.osaka-u.ac.jp/research/351⟩（参照 2021-08-22）．