Évaluation des mesures d'équité pour les systèmes biométriques par génération de biais contrôlés

KN. SANON[1 2] J. DI MANNO[2] T.GERNOT[1] C. CHARRIER[1] C. ROSENBERGER[1]

[1] Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ,
GREYC UMR6072, F-14000 Caen, France
[2] FIME EMEA, 14000 Caen, France

{neily.sanon, tanguy.gernot, christophe.charrier}@unicaen.fr
{christophe.rosenberger}@ensicaen.fr, {joel.dimanno}@fime.com

Résumé

Les biais dans les systèmes biométriques, tels que ceux liés au genre et à l'ethnie, sont des enjeux technologiques et sociétaux cruciaux à considérer. Dans ce travail, nous suivons un scénario de boîte grise avec une transparence limitée, permettant l'accès au score de comparaison biométrique et au seuil de décision. Différentes métriques d'évaluation ont été proposées pour quantifier l'équité d'un système biométrique. Nous proposons, tout d'abord, une méthode pour analyser les biais dans un système équitable et, ensuite, une étude comparative des métriques à l'état de l'art, en nous concentrant sur la corrélation entre les biais et les métriques dans les systèmes de reconnaissance faciale. Dans notre protocole expérimental, nous utilisons différentes bases de visages et des extracteurs de visages avec différentes fonctions de perte. Les résultats expérimentaux permettent d'évaluer la capacité des métriques à quantifier correctement ou non les biais dans les systèmes biométriques.

Mots clefs

Système biométrique, Évaluation de l'équité, Performance.

1 Introduction

Les systèmes biométriques authentifient les individus en utilisant des caractéristiques physiques ou comportementales uniques et sont désormais largement utilisés dans diverses industries. Cependant, leur efficacité et leur équité peuvent être compromises par différents types biais(Dans notre cas, nous considérerons les biais liés au genre).
Cet article porte sur l'évaluation des biais dans les systèmes biométriques, en particulier dans les situations de boîte grise où les mécanismes internes ne sont pas tous accessibles. Notre étude vise à améliorer l'équité et la fiabilité de ces technologies et à établir de nouvelles références pour les évaluations futures dans des contraintes similaires. Les contributions clés incluent :

1. Une revue des métriques existantes pour l'estimation des biais dans les systèmes biométriques.
2. Une méthode pour contrôler les biais après l'extraction des caractéristiques comme vérité terrain pour valider les métriques d'équité.
3. Un protocole expérimental utilisant deux bases de données et systèmes biométriques faciaux avec des fonctions de perte.
4. Des résultats expérimentaux montrant les avantages de la méthodologie proposée pour comparer les métriques d'équité.

L'article est structuré comme suit : la section 2 couvre les bases des systèmes biométriques et des biais, la section 3 examine les recherches sur l'évaluation de l'équité, la section 4 décrit notre protocole expérimental, la section 5 présente les résultats, et enfin, la conclusion aborde la discussion et les perspectives futures.

2 Contexte

2.1 Système biométrique

Un système biométrique vérifie ou identifie un utilisateur en utilisant des caractéristiques uniques. Les modalités typiques incluent les empreintes digitales, le visage, la voix, l'iris, etc. Le système fonctionne suivant plusieurs étapes définies ci-après :

Capture : Le système capture des données biométriques brutes, telles qu'une image du visage de l'utilisateur qui sera ensuite détecté par un détecteur de visage permettant de localiser la zone du visage dans l'image.

Extraction : Les caractéristiques sont extraites de l'échantillon de visage utilisant généralement des réseaux neuronaux convolutifs tels que Inception ou ResNet50.

Comparaison : Les caractéristiques extraites sont comparées à une base de données pour de l'identification ou au modèle biométrique de référence pour de l'authentification. Ceci s'opérant grâce à des calculs de distance telles que le cosinus ou Manhattan, générant ainsi un score.

Décision : Si le score est en dessous (ou au-dessus) d'un seuil fixé, la vérification de l'identité est accordée.

2.2 Définition des biais

Dans cette étude, nous interprétons le biais comme une déviation par rapport à une norme. Il peut se manifester sous différentes formes : statistique (déviations numériques par rapport aux valeurs attendues), moral (déviations par rapport aux normes éthiques), ou encore sous d'autres aspects légaux, sociaux et psychologiques. Nous considérons un système biométrique équitable s'il fonctionne de manière cohérente avec des bases de données biaisées et non biaisées, que ce soit en termes de données démographiques, de condition physique, de qualité de l'image, d'environnement ou d'accessoires. En classification, cela signifie une catégorisation précise indépendamment du biais introduit. Pour notre étude sur l'authentification faciale impliquant le genre, l'équité implique une précision de reconnaissance égale entre ces groupes. Dans la section suivante, nous analysons l'état de l'art dans l'évaluation de l'équité des systèmes biométriques.

3 Etat de l'art

Un ensemble de données biométriques se compose d'utilisateurs $U_i,\ i = [1,\ldots,N]$ (où N est le nombre d'utilisateurs), chacun avec des échantillons biométriques $S_{i,j},\ j = [1,\ldots,M]$ (où M est le nombre d'échantillons par utilisateur). Les utilisateurs appartiennent à des catégories démographiques telles que le genre $d_i = \{\text{masculin}, \text{féminin}\},\ i = [1,\ldots,N]$. Un système biométrique est équitable s'il fonctionne de manière cohérente pour toutes les catégories d'utilisateurs. La performance est évaluée en examinant deux principales erreurs : le taux de fausses correspondances (FMR) et le taux de fausses non-correspondances (FNMR). Les fausses correspondances sont souvent dues à des caractéristiques non uniques, tandis que les fausses non-correspondances peuvent résulter du bruit de l'échantillon, de modèles de mauvaise qualité ou de changements dans les données biométriques au fil du temps. Le seuil de décision τ a un impact significatif sur ces erreurs. L'étude des biais dans les systèmes biométriques est un domaine en pleine croissance, mettant en lumière les défis pour garantir l'équité des performances. En 2021, Drozdowski et al. [1] ont exploré l'impact des caractéristiques démographiques sur la performance de reconnaissance, en soulignant les problèmes entre différents groupes ethniques, genres et âges. Howard et al. dans [2] ont montré que des facteurs environnementaux tels que l'éclairage peuvent introduire des biais dans la reconnaissance faciale.

Différentes approches abordent les biais dans les systèmes biométriques. Schuckers et al. [3] offrent une perspective statistique, en considérant les variations pouvant survenir par hasard (erreur de type I). Fang et al. [4] ont proposé la métrique Accuracy Balanced Fairness (ABF).

Nous avons retenu quatre métriques, à savoir trois sur les résultats différentiels et une sur la performance différentielle, telles que définies ci-après :

1. **Taux de Discrépance d'Équité (FDR)** [5] : Mesure les différences de performance entre les groupes démographiques en comparant le FMR et le FNMR à un seuil donné τ.

$$FDR = 1 - (\alpha \times A(\tau) + (1-\alpha) \times B(\tau)) \quad (1)$$

où

$$A(\tau) = \max_{i,j} \left(\left| FMR^{d_i}(\tau) - FMR^{d_j}(\tau) \right| \right)$$

$$B(\tau) = \max_{i,j} \left(\left| FNMR^{d_i}(\tau) - FNMR^{d_j}(\tau) \right| \right)$$

représentant l'écart maximal de fausses correspondances et de fausses non correspondances suivant les démographies considérées.

2. **Taux d'Inégalité (IR)** [6] : Évalue les disparités en calculant le rapport des valeurs maximales et minimales de FMR et FNMR.

$$IR = \left(\frac{\max_{d_i} FMR(\tau)}{\min_{d_i} FMR(\tau)} \right)^{\alpha} \times \left(\frac{\max_{d_i} FNMR(\tau)}{\min_{d_j} FNMR(\tau)} \right)^{1-\alpha} \quad (2)$$

3. **Taux d'Agrégation de Gini pour l'Équité Biométrique (GARBE)** [7] : Utilise le coefficient de Gini pour mesurer l'équité en agrégeant FMR et FNMR.

$$G_x = \left(\frac{n}{n-1} \right) \left(\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}} \right) \quad (3)$$

$$\text{GARBE}(\tau) = \alpha A(\tau) + (1-\alpha) B(\tau) \quad (4)$$

où $A(\tau)$ and $B(\tau)$ sont les coefficients de Gini pour le FMR et le FNMR respectivement.

4. **Indice de Séparation de l'Équité (SFI)** [8] : Quantifie la capacité à distinguer les scores authentiques des imposteurs entre les groupes démographiques d_i (où $i \in$ [1,D]).

$$SFI_N = 1 - \frac{2}{D} \sum_{i=1}^{D} |z_{Si} - z_{S\text{mean}}| \quad (5)$$

où $z_{Si} = |\mu_{Gi} - \mu_{Ii}|, \quad i = 1, 2, \ldots, D$ et $z_{S\text{mean}} = \frac{1}{D} \sum_{i=1}^{K} z_{Si}$

Ces métriques suivent des approches similaires, et évaluer leur fiabilité est un défi. Ce travail propose de comparer ces métriques en utilisant un protocole rigoureux défini.

4 Protocole expérimental

Dans cette section, nous détaillons les expériences suivies pour comparer les métriques d'équité décrites ci dessus.

4.1 Base de données

Dans ce travail, nous avons utilisé deux ensembles de données publiques de visages. Le premier est LFW10, un sous-ensemble du dataset LFW [9], qui contient 158 sujets avec plus de 10 apparitions. Le second est DemogPairs [10], connu pour son équité en matière de genre et de ces trois ethnicités (Asiatique, Noir, Blanc).

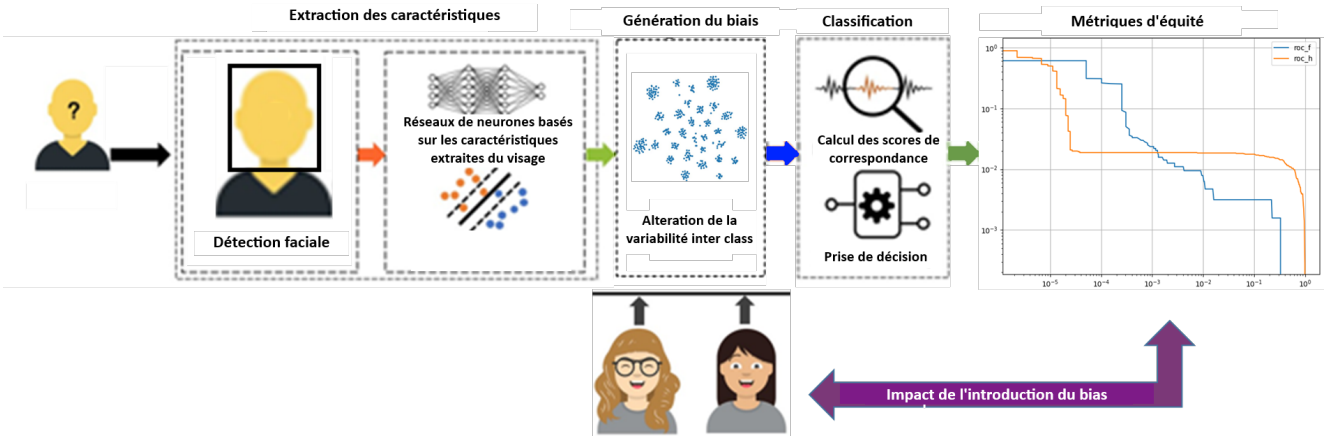

FIGURE 1 – *Méthodologie proposée pour l'évaluation des métriques d'équité.*

4.2 Systèmes biométriques

Nous utilisons MTCNN [11] pour la détection des visages, connu pour sa haute précision. MTCNN est pré-entraîné sur VGGFace, qui se compose de 59,3% de sujets masculins. Pour l'extraction des caractéristiques, nous utilisons les modèles CNN pré-entraînés suivants, chacun étant entraîné avec des fonctions de perte spécifiques pour une extraction précise (512 caractéristiques par modèle) :

Inception-ResNet V1 + Softmax combine l'architecture Inception pour l'extraction multi-échelle de caractéristiques avec ResNet pré-entraîné sur VGGFace2.

InsightFace + ArcFace [12] utilise la perte angulaire additive pour améliorer la précision de la reconnaissance en introduisant une marge angulaire dans la fonction de perte.

4.3 Scénarios de test

Nous avons utilisé deux scénarios : 1 non biaisé et un 2nd biaisé. Initialement, nous avons utilisé des échantillons biométriques sans modification. Ensuite, nous avons ajouté du bruit Gaussien aux caractéristiques pour biaiser le genre femme. Les étapes, résumées dans la Figure 1, sont :

1. **Extraction des caractéristiques et des étiquettes :** Extraire les caractéristiques et les étiquettes des base de données.
2. **Génération de biais :** Ajouter du bruit Gaussien pour biaiser les caractérisques d'une démographie, en occurence les femmes.
3. **Calcul des scores de correspondance :** Déterminer les scores légitimes et imposteurs.
4. **Analyse des taux d'erreurs (FMR et FNMR) :** Calculer les taux de fausses correspondances et de fausses non-correspondances sur différents seuils.
5. **Évaluation des performances :** Appliquer les FNMR et FMR en tenant compte du genre avec des visualisations pour montrer l'impact démographique sur les biais et la précision.
6. **Calcul des métriques d'équité :** Calculer les métriques sur les données non biaisées et biaisées, en contrôlant le biais avec la valeur de bruit σ.
7. **Calcul des corrélations :** Utiliser la corrélation de Pearson pour quantifier comment les métriques d'équité répondent au biais synthétique, en visualisant les relations entre les métriques et le biais.

5 Résultats expérimentaux

Les résultats obtenus avec ce protocole peuvent être divisés en trois parties : la performance initiale des systèmes, le comportement des systèmes biométriques face aux biais, et l'évaluation des métriques corrélant biais et bruit.

5.1 Évaluation des performances

Nous évaluons l'efficacité des deux systèmes biométriques proposées. Nous calculons l'aire sous la courbe (AUC) pour mesurer la capacité du système à minimiser les taux de fausses correspondances et de fausses non-correspondances à travers différents systèmes et bases de données. Une AUC minimale indique une distinction efficace entre différents groupes. Nous avons observé une AUC de 0,02 pour les deux systèmes, suggérant que nos systèmes sont alignés avec la définition de l'équité de la section 2.2. Ensuite, nous examinons les différences de performance entre les catégories démographiques en utilisant la technique de biais synthétique proposée.

5.2 Introduction de biais

L'objectif de cette technique d'altération est de biaiser les systèmes pour évaluer la sensibilité des métriques aux biais. Nous introduisons du bruit avec un σ allant de la moyenne de l'écart-type intra-groupe $\sigma_{\text{std intra group}}$ à quatre fois cette valeur ($\sigma \in [0, 4 \times \sigma_{\text{std intra group}}]$), obtenant ainsi dix valeurs de bruit dans cet intervalle. La moyenne des écarts-type intra-groupe est de 0,039 pour Demogpairs, et de 0,044 pour LFW10. La Figure 2 illustre l'impact de ce bruit sur les femmes, projeté à l'aide de l'algorithme t-Distributed Stochastic Neighbor Embedding (t-SNE).

TABLE 1 – *Corrélation (Valeur absolue en pourcentage) - Genre*

Métrique	FDR					IR					GARBE					SFI
α (si il existe)	0	0.25	0.5	0.75	1	0	0.25	0.5	0.75	1	0	0.25	0.5	0.75	1	
ARCFACE/Demogp	83	83	83	83	80	83	89	**89**	88	80	64	74	76	74	70	83
ARCFACE/LFW10	83	82	82	82	82	83	89	**90**	**90**	83	83	82	81	77	53	81
INCEPTION/Demogp	83	83	83	82	81	83	87	**88**	87	81	78	80	82	83	84	84
INCEPTION/LFW10	84	84	84	82	79	84	**88**	**88**	87	79	80	80	80	80	80	85

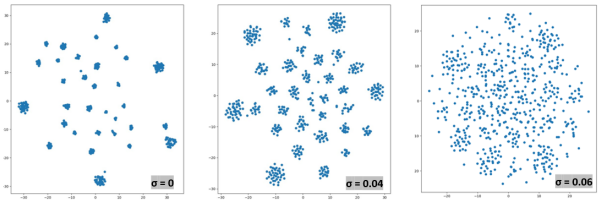

FIGURE 2 – *Introduction de bruit (des caractéristiques initiales aux biais progressifs)*

5.3 Métriques d'équité

Nous évaluons l'équité de ces systèmes biométriques après l'introduction de biais. Les points clés à considérer sont : **1-)** Le biais a été introduit en ajoutant des perturbations aux variations intra-groupe, augmentant le taux de fausses non-correspondances et dégradant l'expérience utilisateur. L'analyse doit se concentrer sur cet aspect. **2-)** La valeur alpha dans FDR, IR et GARBE vise à équilibrer le FMR et le FNMR. Une valeur alpha plus basse est critique lorsque le FNMR est élevé. **3-)** La valeur d'équité maximale pour FDR est 1, tandis que pour IR et GARBE elle est 0. SFI augmente avec l'équité.

Sans biais, Arcface et Inception ResNet semblent équitables pour le genre. Lorsque le biais est introduit, toutes les métriques reflètent un changement dans l'expérience utilisateur, montrant une transition de l'équité à l'iniquité à un certain taux. Les corrélations entre les niveaux de bruit et les valeurs des métriques ont été calculées, en tenant compte de l'aire sous les courbes des métriques en utilisant la règle trapézoïdale.

Le comportement des métriques à travers les datasets et les systèmes est résumé dans le Tableau 1. Demogpairs montre des corrélations stables (80%-89%), indiquant une capture efficace des biais. LFW10, cependant, montre des corrélations variables, surtout pour GARBE avec un α de 0,75 (53%-80%). Cela suggère que les métriques sont moins robustes face à différents types de biais. IR démontre la meilleure corrélation globale. La structure de cette métrique semble intéressante pour une investigation plus approfondie.

6 Conclusion et perspectives

Notre étude des biais dans les systèmes biométriques montre un fort intérêt en raison des exigences strictes de certification en termes de précision et d'équité. Nous comparons les métriques de biais en introduisant des perturbations affectant le taux de fausses non-correspondances. Les métriques semblent stables par rapport aux biais, avec IR en tête. Les recherches futures pourraient ajouter des biais par variation intra-groupe tenant compte à la fois du volet sécurité et expérience utilisateur et développer des mesures qui ne reposent pas sur des paramètres déterministes (comme le paramètre alpha) mais prenant en compte les seuils.Aussi, on pourrait étendre l'étude à d'autres bases de données et extracteurs.

Références

[1] Pawel Drozdowski, Christian Rathgeb, et Christoph Busch. Demographic Fairness in Face Identification : The Watchlist Imbalance Effect, Juin 2021. arXiv :2106.08049 [cs].

[2] John J Howard, Yevgeniy B Sirotin, et Arun R Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. 2019.

[3] Michael Schuckers, Sandip Purnapatra, Kaniz Fatima, Daqing Hou, et Stephanie Schuckers. Statistical Methods for Assessing Differences in False Non-Match Rates Across Demographic Groups, Août 2022.

[4] Meiling Fang, Wufei Yang, Arjan Kuijper, Vitomir Struc, et Naser Damer. Fairness in face presentation attack detection. *Pattern Recognition*, 147 :110002, Mars 2024.

[5] Tiago De Freitas Pereira et Sebastien Marcel. Fairness in Biometrics : A Figure of Merit to Assess Biometric Verification Systems. 4(1), Janvier 2020.

[6] P. Grother. Demographic differentials in face recognition algorithms. EAB Virtual Event Series - Demographic Fairness in Biometric Systems, 2021.

[7] John J. Howard, Eli J. Laird, Rebecca E. Rubin, Yevgeniy B. Sirotin, Jerry L. Tipton, et Arun R. Vemury. Evaluating Proposed Fairness Models for Face Recognition Algorithms. Lecture Notes in Computer Science, Cham, 2023.

[8] Ketan Kotwal et Sebastien Marcel. Fairness Index Measures to Evaluate Bias in Biometric Recognition, Juin 2023.

[9] Gary B Huang, Manu Ramesh, Tamara Berg, et Erik L-M. Labeled Faces in the Wild : A Database for Studying Face Recognition in Unconstrained Environments.

[10] Isabelle Hupont et Carles Fernández. Demogpairs : Quantifying the impact of demographic imbalance in deep face recognition. 2019.

[11] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, et Yu Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. 23(10), Octobre 2016.

[12] Jiankang Deng, Jia Guo, Niannan Xue, et Stefanos Zafeiriou. ArcFace : Additive Angular Margin Loss for Deep Face Recognition. Juin 2019.