Design-driven Deception of Face Recognition: An Empirical Study

PETRELLI, Daniela <http://orcid.org/0000-0003-4103-3565>, DULAKE, Nick <http://orcid.org/0000-0003-1841-5848>, MOLINARI, Gianni <http://orcid.org/0009-0000-5660-3407> and CIRAVEGNA, Fabio <http://orcid.org/0000-0001-5817-4810>

Design-driven Deception of Face Recognition: An Empirical Study

Design-driven Deception of Face Recognition

DANIELA PETRELLI[1]
Department of Design, Politecnico di Milano, Italy, daniela.petrelli@polimi.it

NICK DULAKE
Art Design and Media Research Centre, Sheffield Hallam University, UK, n.dulake@shu.ac.uk

GIANNI MOLINARI
Dipartimento di Informatica, Università of Torino, Italy, gianni.molinari@edu.unito.it

FABIO CIRAVEGNA
Dipartimento di Informatica, Università of Torino, Italy, fabio.ciravegna@unito.it

This paper takes a design-driven human-centred approach to Face Recognition Technology (FRT). In a process of Research through Design we first generated 120 ways to dodge face recognition, then distilled and tested 50 concepts in the lab. The 19 disguises that successfully bypasses FRT informed the implementation of 7 disguises initially tested with 14 white participants walking through a hall, a corridor, a control gate. The control gate led to a larger study (39 participants of different ethnicities) to assess the effectiveness of the disguises in bypassing 3 open-source FR models using 3 different distance metrics and 4 backends. We compare our real-life evaluation of design-generated disguises against previous and current computing research: while maliciously crafted digital perturbation attacks work well, they do not capture the complexity of live FRT opening up opportunities for future research.

CCS CONCEPTS • Human computer interaction (HCI) • Human and societal aspects of security and privacy

Additional Keywords and Phrases: face recognition, public places, privacy, presentation attacks, design, interactive devices, ethnicity, evaluation, dodging face recognition, evading face recognition.

1 INTRODUCTION

Facial recognition is a consolidated yet still growing biometric technology by which the face of a person is recognised as known [Castelvecchi 2020]. The underpinning principle is that the human face is a (unique) combination of features (i.e., eyes, nose, mouth, face contour) that can be used to identify an individual. For its nature, facial recognition is very simple to use and therefore it has become a common biometric authentication system now available on smartphones, home and building security, retailers, border control [Kortli et al. 2020]. Typical scenarios of use are security for the individual, e.g., unlocking personal devices and apps; for companies, e.g., checking people in and out of work; and in public spaces, e.g., to monitor the crowd and rapidly identifying specific individuals [J. Zhang et al. 2021]. When used in public spaces as a way to augment CCTV cameras with the power of AI, facial recognition is labelled by critics as a tool to create a surveillance society where citizens' rights are eroded, and democracy is slowly moving toward authoritarianism [Polyakova and Meserole 2019]. These criticisms become news when face recognition is used by the police to control demonstrations taking place [Sinmaz 2023] or in crime investigations sometimes resulting in individuals being falsely accused [Hill 2020, Hill 2023]. While the use of facial recognition is accepted by the vast majority of citizens for reasons of public security [Bu 2021], more questionable is its use to monitor and regulate people's behaviour, from limiting

[1] Corresponding author.

ACM 1557-7325/2025/10-ART
https://doi.org/10.1145/3769675

shoplifting [Chivers 2019] to school attendance and 'good citizenship' [Bu 2021]. The spreading of the use of facial recognition means companies are pushing ahead with technology development and selling internationally to both governments and individuals [Hawkins 2018]. National and international regulations take time to be developed and agreed therefore critical aspects such as acceptable accuracy, the monetisation of very sensitive data - our faces, privacy rights violations, data collection and use [Bu 2021] are still a matter of debate. Social science literature has since long discussed the ethics of facial recognition [Selinger and Leong 2021] and computer scientists are now aware of the ethical implication of their work [Van Noorden 2020]. Motivated by this societal awareness, some researchers have started to investigate people's attitude and to address existing concerns by introducing privacy-preserving mechanisms as part of their system design [S. Zhang et al. 2021; Löbner et al. 2023]. Nevertheless, computing research still mostly focuses on lab experiments on system performance and technical challenges widening the gap between research and its real application.

Our research looks at Facial Recognition Technology (FRT in the following) deployed in public spaces for the purpose of monitoring and surveillance. It is a very specific scenario where the control of the video recorded, its analysis, use and storage are entirely in the hands of those who provide surveillance services (to the government or the private sector) while the people whose face is captured do not have any power to prevent being recorded and may even not be aware FRT is in operation. In this complex and articulated landscape, we take a design-driven human-centred perspective of the sociotechnical system of technology, people, and places. While most research in this area starts from the knowledge of how FRT works and devises computational ways to evade it [Sharif et al. 2016; Zhou et al. 2018; Pautov et al. 2019; Zhu et al. 2019; Zolfi et al. 2021; Vakhshiteh et al. 2021, Jaiswal et al. 2022, Li et al. 2023], our research is computing-agnostic and founded on the human question "if I did not want to be recognised, what would I do?" We started with an intense design phase: of an initial set of 120 concepts[2] 50 were fast-prototyped and tested by the researchers in the lab (Section 4). The 19 concepts that successfully bypassed FRT were implemented as 7 disguises (4 static and 3 interactive) and tested with 14 white participants walking through 3 cameras positioned on a corridor door, on the corner-ceiling of a hall, and in a frontal position as in a controlled access gate (Section 5). The FRT performed poorly in the corridor and hall. A final more extended experiment focussed on the access control setting with 39 participants from different ethnicities. Our findings are not always aligned with existing literature pushing us to reflect on why this is the case (Section 6).

2 DEFINING THE RESEARCH CONTEXT AND ITS ORIGINAL CONTRIBUTION

The scenario of reference for our research is FRT deployed in public indoor places where live cameras are used for monitoring and surveillance. This focussed goal brings some specificities that distinguish our investigation and define our original contribution respect to other FRT research conducted in the same or close contexts.

We started with a generative Research through Design phase to conceptualise, test and select human-centred proof-of-concepts that could be made by, for example, human-right activists against FRT. The promising concepts were then prototyped as physical disguises to be tested first in the lab and then in realistic experiments. The process of creating the disguises did not consider the capabilities of the FRT, rather it started from what professional designers imagined could be effective ways to go undetected. Much previous research, instead, implements disguises that exploit known FRT weaknesses (e.g. bespoke adversarial patches [Pautov et al. 2019]; sharp light change [Li et al 2023]) requiring expertise in computer science, some knowledge of the FRT being targeted, the camera and sensors in use [Bisogni et al 2021]. Some of the concepts we generated share some features as those in the literature (a pair of glasses [Sharif et al. 2016], face projection [Li et al. 2023], makeup [Zhu et al. 2019] or a hat [Zhou et al. 2018]). However, our disguises do not require any computing knowledge (such as perturbation algorithms to feed obfuscation systems [Rosenberg et al. 2023, Vakhshiteh et al. 2021]) and could be made with minimal DIY skills. Therefore, our work takes a human-centred perspective, complementing, expanding and challenging existing research by offering similar, yet different physical disguises generated following a design approach. Some of our results contradict papers in the literature pointing

[2] 'Concept' is a term commonly used in design to indicate a rough idea of a potential solution to a problem or a question. A concept does not have to be realistic or feasible, it functions as an inspiration and as a starting point.

to key role of computational deceptions such as gradient-based perturbation rather than the physical means of delivering it such as a mask, a pair of glasses or a makeup, as discussed in Section 7.

Generative design does not pursue a single solution, rather it explores the problem through multiple and diverse options simultaneously [Brown 2009], in our case physical disguises with different features. Following a design thinking approach we also quickly prototyped and tested many concepts in the lab to help us select the most promising to be taken to implementation and testing. Existing research mostly focusses on a single object type (e.g. mask [Zolfi et al. 2021], glasses [Sharif et al. 2016], hat [Zhou et al. 2018]). Instead, our informal tests in the lab and more structured realistic evaluations simultaneously compare multiple different types of wearable disguises trying to assess the most effective ones. This comparison across disguises is a unique contribution.

During our study we conducted three different assessments of FRT against the disguises, always using live video. The first was a lab test (Section 4.1): the researchers presented themselves to the FRT wearing one of the 50 proof-of-concepts to check if they were recognised or not. The result of this test informed the prototyping of the 7 disguises, 3 of which were interactive. In the second test, 14 white participants walked along a corridor, around a hall, and in front of a camera wearing the disguises; in the third experiment we expanded the number of participants to 39 including white, brown and black people. We also tried to balance the sample by inviting participants of different ages and different gender. This quasi-realistic setup makes our research unique as the data used to test FRT in the lab is generally a database of images computationally manipulated [Hernandez-Ortega et al. 2019] or images captured in ideal conditions with the subject and the camera in fixed positions within a controlled environment and perfect illumination (e.g., [Sharif et al. 2016, Li et al. 2023]). To the contrary, our participants were wearing the disguises and walked through three different spaces with different illuminations and the camera placed at different angles. Finally, we run multiple tests to compare different FRT configurations over a period of 3 years, from Spring 2021 to Summer 2024 showing how FRT has improved over time and providing empirical evidence of the best performing configuration. Therefore, our research is the first to show the combined effect of free movement of people, ethnicity, disguises, camera angle, and light variation on the performance of multiple FRT systems over time. We also performed a series of cross tests to assure our results are consistent and reliable, including: the test of different video samples rating to balance accuracy and processing times; logistic regression to evaluate the impact of each variable (disguise, gender, ethnicity) on the predictive capacity of one of the top-performing FRT pipelines; open-world settings with larger databases of identities to check if the performance changes as the number of individuals increases. These tests could be a guide for other researchers.

3 FACE RECOGNITION TECHNOLOGY

In this section we review the literature on FRT to identify individuals and deception strategies to avoid identification in video streams. We attempt to include both academic and commercial systems, although public reports on real uses of commercial systems are very rare. In reviewing the literature, we limit our effort to the specific scenario of reference: FRT deployed in public indoor places where live cameras are used for monitoring and surveillance where the people whose face is captured do not have any power other than to wear a disguise. Therefore, user-controlled privacy [Aditya al. 2016], devices that guarantee privacy by adding noise to the image files [Vakhshiteh et al. 2021, Jaiswal et al. 2022], by "cancelling" bystanders [Hasan et al. 2020], or camera that shut when people are in view [Steil et al 2019, Aditya et al. 2016] are excluded from this review because those pursue the opposite goal. Mechanisms that identify people by other means than only their faces (e.g. clothes and pose [Oh et al. 2016], gait [Wang et al. 2018], information in Linkdin or FaceBook [Acquisti 2014]) are also considered out of scope.

3.1 Face recognition technology

Face recognition is a biometric tool to identify faces in images, both photos and video stills. Although there is a tendence to think of FRT as pure software, the starting point is the sensing of the world through a camera (that captures the image), possibly complemented by additional sensors that can help to better interpret the scene, for example an infrared camera, a depth camera, or an eye tracking sensor [Kortli et al. 2020]. The performance of FRT depends on the accuracy of the sensing system and additional sensors to a camera could

improve performance in conditions of illumination variation, or additional sensors could prevent 'presentation attacks' [Hernandez-Ortega et al. 2019]. The processing pipeline is articulated in several steps some of which are common across different purposes (see [Kortli et al. 2020] for recognition techniques):

1. *Face detection* scans the image seeking faces [Kortli et al. 2020]. The output is a bounding box for each face detected in the image. Face detection may fail because of variation in the illumination of the scene, partial obstruction of the face and facial expressions.
2. *Face alignment* locates face landmarks such as eyes, nose and mouth and aligns them generating a frontal and well-lit view of the face [Castelvecchi 2020].
3. *Features extraction* extracts the geometry of the face(s) to create a 'faceprint', that is unique for every person, and encodes the faceprint into the feature space [Kortli et al. 2020]. The output of features extraction can be used in multiple ways, for example in *Facial Analysis* to classify a person respect to their age or gender or to assess the person's emotional state [Castelvecchi 2020], or to populate the database of identities used in *Facial Identification* [Kortli et al. 2020].
4. *Face matching* confirms or checks the identity of a person. *Verification* is a one-to-one comparison in which the faceprint extracted from the image is compared with the one stored. Examples are unlocking a smartphone, passport control or work attendance [Castelvecchi 2020]. *Identification* is a one-to-many comparison in which the faceprint extracted is checked against a database of known people to discover the identity of the person detected [Castelvecchi 2020]. An example is the scanning of a crowd to find people of interest in police operations [Fussey and Murray 2019, Mansfield 2023].

This paper focuses on facial recognition for identification. FRT failure in identifying a person can occur at different points: *detection failure* when no face is detected in the image or a face is detected where there is none; *recognition failure* when a known face is not recognised in the image or the person is recognised as someone else. The rate of failures determines the accuracy of the FRT. The performance of FRT in ideal, controlled environments can be very high: [Kortli et al. 2020] compared different techniques and reported scores of 96-99% with perfect illumination, a fixed pose, optimal computational power, and computational time. Similarly, a comparison of commercial[3] FRTs on a database of 80 celebrities for which excellent visual material is available, e.g., multiple photographs from different angles, found 93-99% accuracy [Raji et al. 2019].

Reliable data of FRT performance in uncontrolled, in-the-wild, real-life use is an exception. Hindsight from a live facial recognition trial carried out by the police in the UK in 2018 showed a correct match of 19.05% (8 correct out of 42 matches found by the FRT used) [Fussey and Murrey 2019]. However, a recent and larger study from 2022 carried out by the police in the UK using real-life recordings in the streets of London in a sunny summer day (i.e., optimal illumination) with a controlled cohort of subjects found a positive identification of 89% across all ethnicities [Mansfield 2023]. It is an impressive increase in performance and a steep drop in the error for false positive (when a person is wrongly identified as being in the watchlist)[4]. Even with a minimal error rate, false positives are particularly critical when the scale of reference is the entire population of a country given the consequences of false accusation on the life of the individuals wrongly accused. As the media report cases of false accusations following FRT misidentifications (by USA police [Hill 2020, Hill 2023, Williams 2020] and by Apple in their shops [BBC 2019]), the accuracy in real-life use and across different FR systems becomes crucial. In addition, while governments are likely to have rules and procedures to safeguard their citizens (e.g. the UK requires a final verification by a police officer before any action is taken [Fussey and Murrey 2019]), the private sector may be acting in a vacuum of regulation rising concerns about potential misuses and abuses (as in recent cases in the USA [Murphy 2023, Bhuiyan 2023, Bhuiyan 2024]).

[3] The largest comparative continuous assessment of FRT commercial algorithms is carried out by the USA National Institute of Standards and Technology NIST using very large datasets of images: companies apply to NIST to have their software tested in both facial identification and analysis. The last report from 2020 is available at https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

[4] [Mansfield 2023] reports an estimated false positive error rate of at 0.017% for a watchlist set of 10,000 and a false positive rate of 0.002% for a set of 1,000.

3.2 Presentation and adversarial attacks

Even if the first successful face recognition system dates back to the 1960s, it has been only with the last advancement of deep learning and the availability of abundant data that FRT has reached accurate performance [Vakhshiteh et al. 2021]. However, deep learning face recognition systems are vulnerable to data variation [Vakhshiteh et al. 2021]. Thus, much effort is spent on making the technology more robust to attacks [Hernandez-Ortega et al. 2019].

The literature distinguishes *physical attacks*, those modifying the appearance of a face before image capturing, and *digital attacks*, those modifying the captured image by adding 'noise' (or perturbation) to create an adversarial image that deceives FRT but still shows the original face to a human eye [Vakhshiteh et al. 2021]. In essence, digital attacks exploit known weaknesses of FRT to mislead it.

Our research investigates physical attacks in the real world, called *presentation* or spoofing *attacks*. Such attacks on facial sensors can be carried out by simply presenting the artifact to the camera and do not require any knowledge on how the biometric system works [Ramachandra and Busch 2017]. A presentation attack can have different purposes: the 'active impostor' wants to be identified as a specific individual (*impersonation*); the 'active impostor' aims to be identified as any other individual (*dodging*); and the 'concealer' intends to evade being recognised as any individual known to the system (*evasion*) [Ramachandra and Busch 2017, Vakhshiteh et al. 2021]. Impersonation attacks make use of artifacts such as a photo, video or 3D mask of the target person. Software mechanisms have been implemented in FRT to detect the 'liveliness' in the face (e.g., skin texture and blink detection against images and 3D masks) or by measuring the depth of the scene through reflectance (e.g., variation of camera focus or infrared against the use of video on tablet) [Ramachandra and Busch 2017, Hernandez-Ortega et al. 2019].

Impersonation can be also achieved using 'adversarial patches', i.e., print out of perturbations to be worn on glasses [Sharif et al. 2019], on a t-shirt [Xu et al. 2020], disposable masks [Zolfi et al. 2021], or on part of the face [Pautov et al. 2019]. These patches are created ad-hoc to impersonate a specific person, a result also achieved by projecting perturbations on the face [Shen et al. 2019, Nguyen et al. 2020, Li et al. 2023]. However, the use of adversarial perturbation in presentation attacks requires substantial computing knowledge to create patches for impersonation out of an image of the target person, therefore it cannot be considered an attack easy to carry out.

In this paper we explore different forms of presentation attacks and test if the disguises can dodge or evade FRT. We implemented and tested several strategies similar to a few in the literature, yet our work is unique in three ways: (1) we implement wearable prototypes, (2) in the evaluation we asked participants to wear the disguises and walk towards a camera, and (3) we compare the effectiveness of the disguises in bypassing FRT. Literature on physical attacks very rarely presents prototypes used in real life conditions. For example, attacks by projection on the face [Shen et al. 2019, Nguyen et al. 2020, Li et al. 2023] place the projector far from the person who stands still while their image is captured; conversely, we explored and implemented projection as part of a hat (Figures 9, 10, 11) that participants wore while walking around during the evaluation. Another example is the use of makeup in an adversarial attack that automatically add eye and lips makeup to images or a video feed [Lin et al. 2022]: we experimented with different makeups by painting our faces and found different results. Section 3 discusses our design with reference to examples in the literature to highlight when our approach differs and when it is similar.

4 DESIGN RESEARCH TO EXPLORE FACE RECOGNITION DECEPTION

This section discusses the design process. We used a Research through Design methodology [Stappers 2007, Koskinen et al. 2011] starting with a phase where the brief of bypassing FRT is expanded and explored from different angles. The concepts generated are then assessed in a convergent phase, some are eliminated, other are aggregated. The most promising are taken forward to a phase of experimentation with material, fabrication and testing that is instrumental to build an understanding of what could work and what could not. On the bases of empirical evidence a few concepts are then developed to be evaluated in a more rigorous way with recruited participants.

4.1 Proof of concepts

As discussed in 3.2, attacks to FRT exploit known weaknesses and generally require computer science expertise to implement the precise attack that will fool FRT. Instead, a human-centred, naturalistic approach runs through our research. We started with the question: "if I did not want to be recognised, what would I do?" and speculate on possible scenario of use, e.g., disguises that would go unnoticed in the street and make the wearer invisible to FRT at will.

Six (6) experienced professional product designers participated in a 2-hour generative workshop. The design brief was to imagine a wearable that could successfully bypass FRT. What the technology could or could not do was not discussed, rather it was left to the intuition of the designers to imagine what could be effective to go undetected. The product designers then proposed concepts explaining what the deception was and how it worked. We stayed open to both simpler and complex concepts irrespective of how 'implementable' or 'realistic' such concepts might appear at first sight (Fig. 1) for their potential to inspire design. Overall, 120 different concepts were generated, several described an interactive piece that could be switched on and off. We did not brief the product designers on which physical attacks have been tried in the scientific community, yet a few of the concepts generated in the workshop resemble presentation attacks tested by computer scientists: 3D mask [Ramachandra and Busch 2017, Hernandez-Ortega et al. 2019], adversarial face mask [Zolfi et al 2021], different types of projection on the face [Shen et al. 2019, Nguyen et al. 2020, Li et al. 2023] and dazzled glasses [Sharif et al. 2016, Sharif et al. 2019]. The key difference between the concepts generated by the designers and those in the literature was the mechanism to bypass FRT, for example the dazzled glasses use 'perturbation' for impersonation, what looks like random bright-coloured patterns placed on the glasses frame [Sharif et al. 2016] while the glasses imagined by the designers generated a hologram when clicked together and were later implemented with a hidden mechanisms for light projection against the camera (Fig. 6).

The workshop was a divergent, generative phase that was followed by a selective, convergent phase [Brown 2009]: the 120 concepts generated in the workshop were critically assessed by the researchers, those similar were aggregated, and 50 were taken forward. Fig. 2 and Fig. 6 show examples of aggregated concepts. The proof-of-concept eye-mask in Figure 2-bottom merges two concepts: a textured mask and a lighting mask in Fig.2-top. Figure 6 shows the merging of two different objects a pair of hearings emitting light and a pair of holographic glasses combined into a proof-of-concept of glasses projecting light. The aggregation of concepts could occur at any time during the design process: Fig. 10 shows how the projection in, already an aggregate of several projection concepts, was implemented with a geometric pattern that was successfully tested as makeup, bottom left in Fig. 3.

Most concepts were quickly mocked-up as proof-of-concept and tested by the researchers themselves in Spring 2021 during COVID lockdown (Fig. 3). The interactive disguises, discussed in Section 4.2.2 and 4.2.3, where warn by the researcher and switched on and off: if the disguise was effective, the FRT system would flip between recognising when the switch was off and missing to recognise the person when the switch was on, as shown in Fig. 10. A few concepts could not be prototyped due to difficulties to procure special materials and access fabrication machinery while in lockdown; however, elements of those concepts were transferred, for example the thermochromic face cover (Fig. 1, centre bottom) shares elements of the 'prosthetic patches' and the 'projection in' concepts, both developed as proof-of-concept (see Fig. 3). In this way we were able to test features commons across multiple concepts. An example is given in Fig. 2 that combines two concepts of mask, one with a textured surface, the other with a light diffuser. Even though this specific exploration was not taken forward to become a prototype to be tested with participants, the process of designing and making it increased our knowledge of material-digital hybrids.

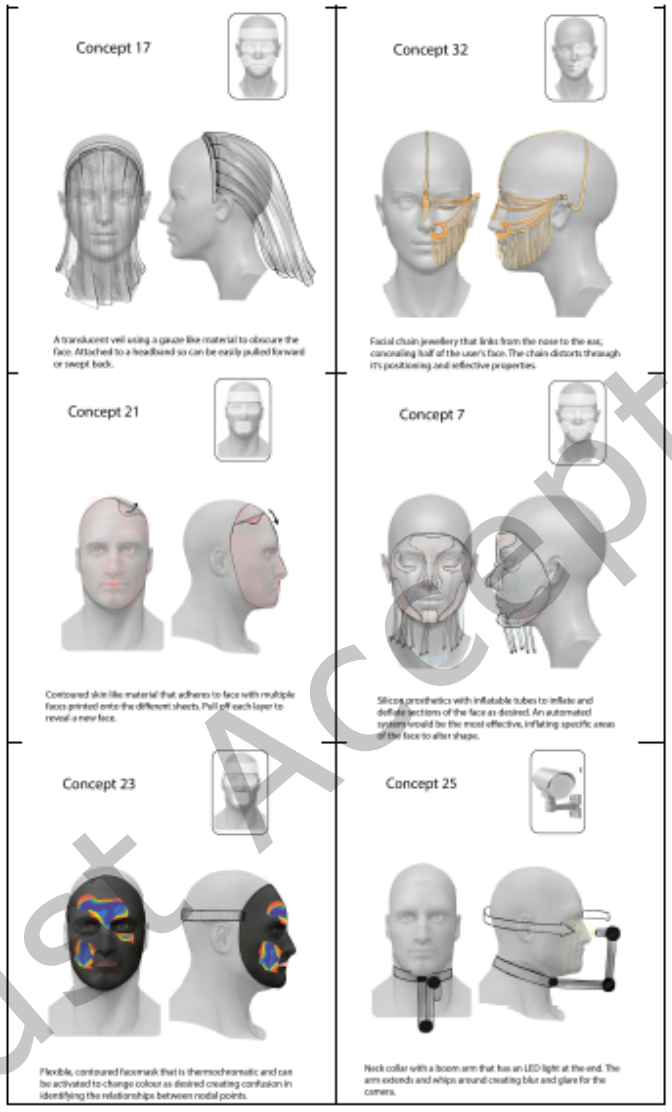


Figure 1. Some of the concepts generated in the workshop, from simple (top) to complex (bottom).

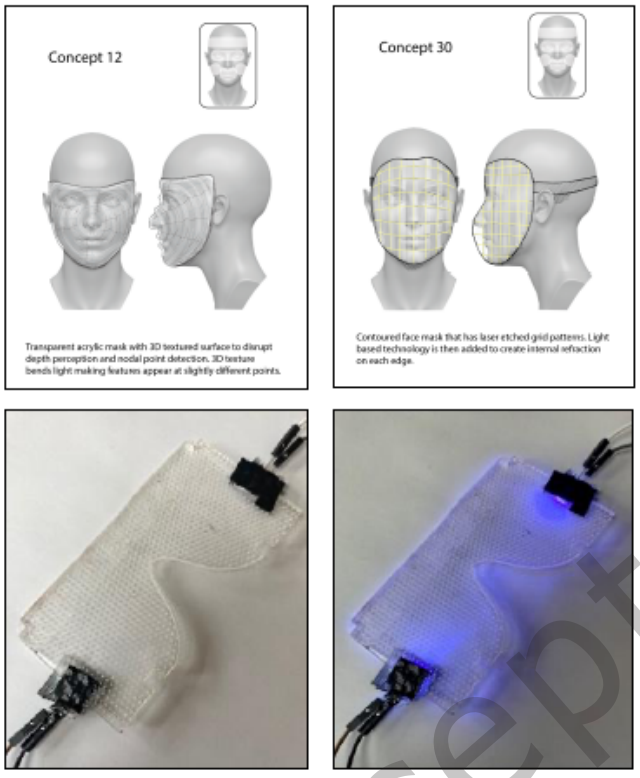


Figure 2 A proof-of-concept (bottom) that combines two similar concepts (top): a mask with a 3D texture (Concept 12) and refracting light (Concept 30). It combines light and mask using internal refraction to create a glow effect. The 3D printed sheet adds the pattern.

The proof-of-concepts were tested against DeepFace[5], running VGG and Facenet models, paired with a laptop or an external camera in Spring 2021. The multiple hardware and software settings were instrumental in the process of gaining an empirical understanding of the variables that come into play. For example, the optics of different cameras may impact the performance of FRT and different models are likely to have different performance: we considered successful only those proof-of-concept that bypassed all configurations. Testing the 50 concepts, or part of their features, was instrumental to build an understanding on which strategies could be effective in deceiving FRT and which were not. Indeed, while face recognition software is available opensource for everybody to use, which are the key facial features for each model is not clearly disclosed. From our experiments, some of which are shown in Figure 3, we inferred that the eyes and the nose bridge were key points for recognition whereas the mouth and the chin were not critical. For example, the masks and the prosthetics leave the eyes and the nose-bridge visible and failed to bypass FRT while all the concepts that successfully bypassed FRT hid the eyes and the nose-bridge.

The tests of the proof-of-concepts eliminated those that did not bypass FRT. For example, most makeups were not effective, possibly because makeup emphasises facial features rather than disguising them, whereas many snoods[6] bypassed FRT successfully, and projection in gave inconsistent results (see Fig. 3 for examples). Interestingly, our tests were not always consistent with the literature, for example the synthetic makeup in [Lin et al. 2022] was effective in deceiving FRT while two of our makeups were detected while one was successful, this in line with [Chen et al 2017] who recorded equally patchy results. Our results also contradict [Shen et al. 2019] face-on-face projection and [Zolfi et al. 2022] nose-mouth mask, see discussion in Section 7.

[5] DeepFace software was downloaded from: https://github.com/serengil/deepface

[6] A 'snood' is a wide ring of knitted material worn as a hood on the head or as a scarf around the neck. When worn around the neck a snood can be easily pulled up to cover the face, as in the concepts explored here.

Figure 3. Some examples of disguises that failed to bypass FRT (top) and succeeded (bottom).

By systematic testing, we identified 19 successful deceptions, organised into four categories classified respect to key features, specifically how easy they were to make (simple vs. complex) and if they were analogue or digital. These four categories of deceptions were the foundation for the second phase of the research, the development of the simplest and the most complex concepts (from prototypes to products) and their assessment under near-real life conditions.

4.2 Prototyping disguises

From the classification of successful disguises, we selected three generic concepts to be implemented in different forms: camouflage, projection out and projection in. At this point we started to consider realistic scenarios of both making and use. The disguises are described below starting from the original concept (in the set of 50) through to experiments with materials and proof-of-concepts (tested by the researchers) to the final implementation ready to be used in the comparative evaluation discussed in section 5. In this process, from early concept to final implementation, we had, at times, to simplify the original concept to make it feasible and implementable or to make it safe and easy for participants to wear during the comparative evaluation. In hindsight, our choice to prefer ease of use and the comfort of the participants taking part in the experiment against maintaining the successful technical solution as tested in the lab drastically compromised the effectiveness of the disguises, as discussed in Section 5 and 7.

Figure 4. Extensive testing: (top) different fabric and invisible bandages; (bottom) different patterns draws and printed.

4.2.1 Camouflage

The concept of camouflage includes a wide range of 'analogue' disguises to wear. A camouflage can be 'long-lasting' such as an elaborate makeup that needs time to be put on and off, or it can be a 'quick' disguise such

as a patterned snood, a face shield or a veil that can be put up and down quickly. The snoods successfully bypassed FRT and the concept was taken further (Fig. 3). The next design iteration of the snoods explored materials, patterns and how to make them (Figure 4). Inspired by the idea of multiple skins to peel off (Fig. 1, Concept 21), we experimented with invisible bandages (also called 'second skin'), an adhesive film applied directly over the skin, for example, to heal burns. The effect on the face is similar to the snood (see Fig. 5) and has the advantage of being quick to dispose of, a key factor in eliminating evidence of a disguise (Fig. 4, top right). Different patterns were designed and tested (Fig. 5, bottom row) and 3 were selected for production (Fig. 5).

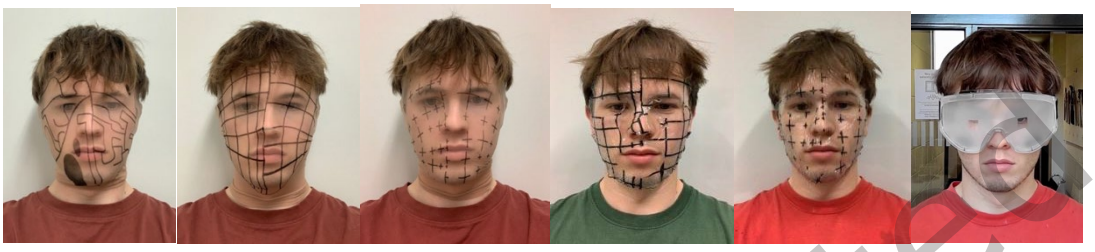


Figure 5. Camouflage: 3 snoods with different patterns printed on tights (left), 2 patterned second skin (middle), goggles (right).

A further simple camouflage was introduced following the observation by a design researcher that they were unable to unlock their phone (that use a one-to-one FRT) when wearing dusty, safety goggles. We therefore introduced this disguise imagining that someone dressed as a maintenance worker could easily walk in the street unnoticed.

4.2.2 *Projection out*

A small set of the 50 concepts proposed a camera-attack strategy: a beam, controlled by the user, is directed toward the camera with the intent of confounding its optics and prevent face detection. These concepts of attacking the camera fed an initial exploration of materials that could reflect infrareds (Fig. 6, middle) as infrared sensors could be added to a standard camera to sense the scene at night [Kortli et al. 2020]. This very quick and effective proof-of-concept fed the exploration of two interactive disguises that used non-visible wavelength light: ultraviolet (UV) led and infrared (IR) led (Fig. 6, third and second from right). The UV interactive glasses were very effective in bypassing FRT under different light conditions whereas the IR glasses failed in daylight conditions with one of the cameras.

Although successful in bypassing FRT, the 8 UV LEDs mounted on the frame made the shape bulky and the wires were unpractical for the evaluation with participants. As we aimed at wearables that participants could easily put on and off, we decided to make the projecting out glasses using an off-the-shelf frame. To eliminate the wires the number of light points was reduced and optical fibres along the top of the frame were used to form a line of 4 light points (Fig. 6, right). The light projected is visible light 400 – 780nm and the battery needed to switch the light on/off is small (UV and IR would have required a much larger battery pack). We were aware that the change from UV/IR to visible light and the reduced points could impact the power of deception of this disguise; what we gained was to move closer to a standard pair of glasses that could be used in real life and observe its wearability and ease of use in the evaluation. If this disguise would be partially successful in bypassing FRT in the evaluation, then the effort to micro-engineer the projection-out glasses to reduce the size of the electronics would be justified.

Figure 6. The concepts of morphing glasses and a camera-attack mechanism that inspired the design of glasses projecting out.

4.2.3 *Projection in*

A set of concepts from the workshop propose hats or headgears that project images or patterns on the face (Fig. 7). Rather than attacking the camera directly, these projections disguise the person by confounding their facial features. It was a promising intuition, and three quick tests were done early in the research, two were successful (Fig. 3): one floods the face with white light, the other projects geometric patterns. This started an exploration of two different headgears, tested as proof-of-concepts. The 'halo hat' is a round brim straw hat with a ribbon of white LED on the bottom of the brim; a switch controls when it is on-off (Fig. 8). It was effective in bypassing FRT, wearable and easy to use thus ready for the evaluation.

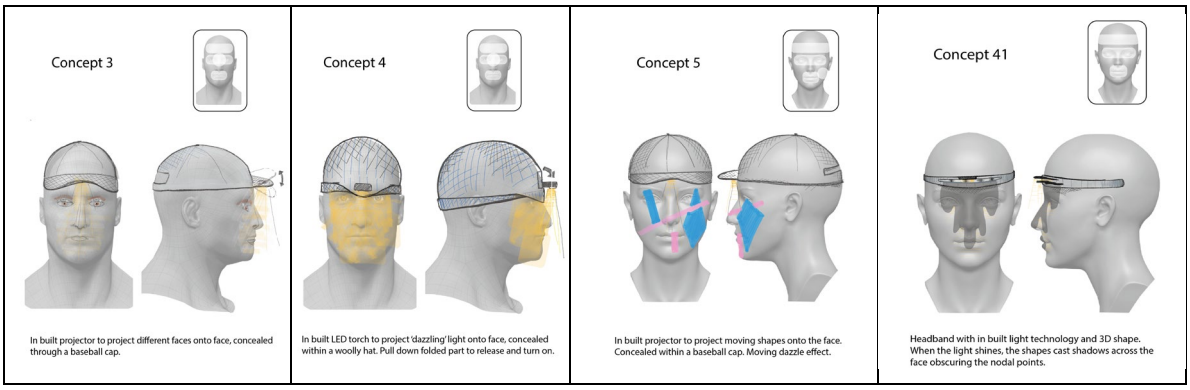


Figure 7. Headgear that 'project in': (1) the face of someone else – tested as proof-of-concept, it failed, see Fig. 3; (2) a dazzling light – the 'halo hat' prototype; (3) geometric shapes tested as makeup succeeded, see fig. 9; (4) shades – tested as the 'rainbow hat' prototype.

Figure 8. The 'halo hat' with a ring of LED around the brim. The on-off control is a simple switch.

More articulated was the exploration of the geometric projection. To become wearable, the pico-projector used in the quick tests in Fig.3 had to be fixed to the head to guarantee an alignment between the position of the head and the projection. A test rig was made using carbon fibre and a bespoke 3D printed holding piece for the pico-projector to be mounted on a bike helmet (Fig. 9, left). In the proof-of-concept the face-projection alignment was consistent and the bypassing of the FRT very effective (Fig. 9, right). Incidentally, the geometric projection implements the geometric makeup, the only makeup to effectively bypass FRT in the test (Fig. 3), another example of merging different features while progressing along the design thinking process [Brown 2009].

Figure 9. The 'projection in' of a geometric patter: FRT recognizes the user when the projection is off but not when it is switched on.

To use this concept in the evaluation we needed to improve its wearability and ease of use. However, this would require much engineering work (e.g., to miniaturise the electronics) that was not affordable in terms of time and expertise. Thus, a much simpler projection-in prototype was implemented for the evaluation (Fig. 10). As for the projection-out glasses, we had to compromise the original concept for the evaluation. The 'rainbow hat' prototype projects a carousel of coloured lights on the face; it is controlled by a smart watch to avoid attracting attention while quickly switching it on and off.

As for the glasses (see 4.2.2 Fig.6), the constraints for disguises that were easily wearable to be used in a naturalistic evaluation by participants external to the project walking around spaces, changed part of the concept, specifically the type of the pattern projected. Given the result of the lab tests (Fig. 3), in changing the pattern we were aware of the potential failure (i.e. to be detected): our intent was to see the result of the evaluation, learn from it, and then review the design, see section 7 Discussion.

Figure 10. The rainbow hat and the smart watch used to control it: a line of coloured led below the visor of the cap projects animated coloured lights on the face. A smart watch controls the on-off switch of the light without attracting attention.

5 UNPACKING FRT PERFORMANCE

5.1 The ethics of evaluating FRT with participants

Conducting research on FRT requires a degree of sensitivity to the subject that only recently has been developing within the computer science community [Von Noorden 2020]. As HCI researchers we were aware of the ethical concerns since the inception of the project and ethical considerations run throughout the whole research, from designing comfortable and safe disguises; to wearability and ease of use to perform the evaluation; from system configuration for assure privacy to personal data dissemination in publications. The evaluation plan was reviewed and accepted by Anonym ethical committee that implements rigorous standard of ethical research[7]. In this section we describe how we delt with ethical issues at different points in the research process.

In moving from concepts tested in the lab by the researchers on themselves to implemented disguises to be worn by participants in the evaluation we chose comfort and ease of use above technical coherence. In this process some disguises were: eliminated because they were too intrusive (the 'second skin' bandage Fig. 5, second and third from the right); redesigned to improve wearability (the UV or IR glasses needed many wires around the head and a bulky power battery to carry around) or to guarantee health and safety (in projection-in, 'halo-hat' in Fig. 8 floods light from above had to avoid the eyes; 'rainbow-hat' in Fig. 10 projects mild LED light on the face). We were aware these choices could lower the effectiveness of the disguise, yet we preferred to have lower results and then to go back to the drawing board with informed knowledge. To guarantee hygiene, we made a large set of one-use snoods and dispose of them after use.

Our commitment to guarantee the privacy of our participants impacted on the technical setup. To assure the recorded videos were always under our control in full, we limited ourselves to open source non-commercial FRT software to be installed on the local machine together with the videos. This was clearly stated in the participants' consent form: it was an essential and indispensable condition of the experiment. We were fully aware that commercial systems offer better performances, yet we intended to use only a configuration that guaranteed participants' facial features stayed local and under our control even if we had to renounce to optimal performance. We are also aware that our videos would be an important resource for other researchers, yet, we committed not to distribute it to maintain the privacy of our participants.

On arrival participants were asked to read the information sheet and sign the consent form which had different options: they could opt not to wear a specific disguise if they felt uncomfortable (a person preferred

[7]A link to the university policy will be added upon acceptance.

not to wear the snoods) or they could opt out of having their faces used in publications or presentations (three participants opted out).

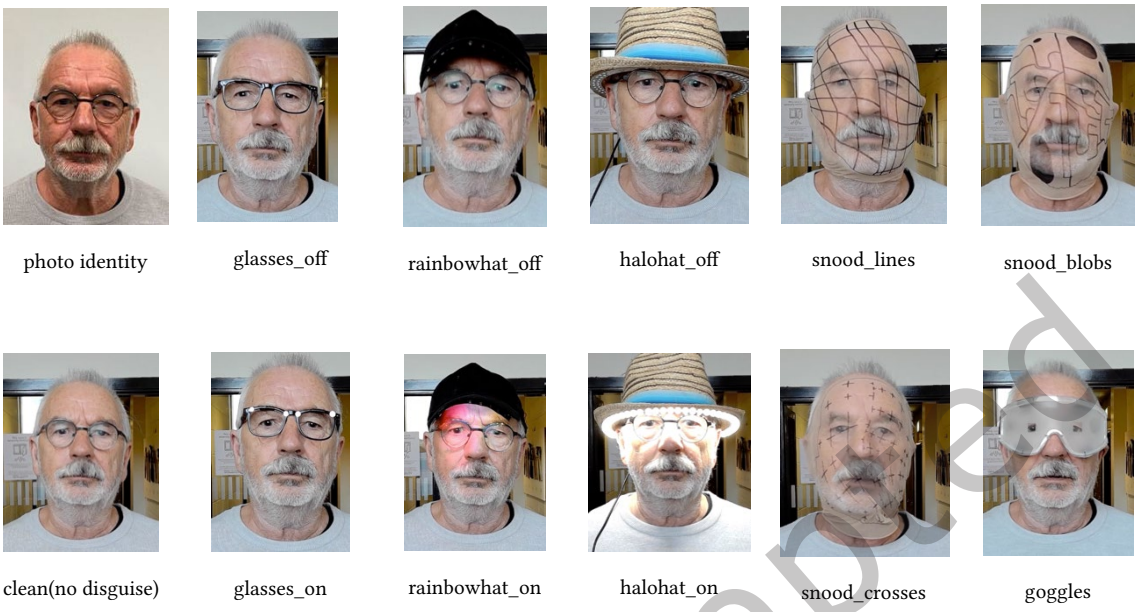


Figure 11. One of the participants wearing the disguises. Top left: the photo portrait used to create the database of identities. Bottom left: a frame from the clean video used as benchmark. From second left the 10 disguises, the interactive are shown as off and on.

5.2 Experimental design and procedure

The evaluation was designed to assess which disguises bypassed FRT in realistic situations while maintaining data confidentiality. We balanced the rigour of a controlled in-lab experiment with a veridic naturalistic setup: in a within-subject experimental design each participant wore all the disguises in turn and walked toward the cameras positioned as to resemble real FR conditions. The experimental design was planned and piloted, changes and amendments implemented. The first experimental variable is the disguise. Of the prototypes developed, the 'second skin' was tested only by 2 researchers as it was considered too intrusive for the participants to wear; however, patterns printed on the second skin were tested with participants in the 3 snood conditions. The three interactive disguises (glasses, halo-hat, rainbow-hat) were evaluated in two states, on and off. The benchmark is 'no disguises', so a clean face. Overall, there are 11 values for the experimental variable *disguise: clean, snood-blobs, snood-lines, snood-crosses, goggles, glasses-off, glasses-on, halo-off, halo-on, rainbow-off, rainbow-on* (Fig. 11).

A proved critical element in FRT performance is the skin colour [Buolamwini and Gebru 2018, Rosenberg et al. 2023]. We recruited 39 participants and distinguish them by skin colour: 22 white participants from both North and South Europe and the Middle East (13 males and 9 females aged 23 to 76); 9 brown participants from India (4 mans and 5 females aged 26 to 32); 8 black participants from Nigeria (4 males and 4 females aged 25 to 35). The sample is too small to test intersectional samples of skin-gender or skin-age, thus we limit the second variable to *ethnicity: white, brown, black.*

Further variables define the *environment*, i.e., the position of the camera with respect to the person to be identified, the FRT configuration, i.e., the *model* (e.g. ArcFace), the *backend* (e.g. OpenCV), and the *distance metrics* used to compare the face identified with those in the database (e.g. Euclidean). These variables changed from the first explorative evaluation to the second more extended one and are discussed in 5.4 FRT performance evaluation.

To reflect a real-life situation where images from passports, identity cards, driving licences, or mugshots database are used to identify an unknow person, we created the database of identities with a single high-quality image for each participant taken full front on a white background at the start of the evaluation (Fig. 11 top left).

To replicate reality as close as possible, we intended to use CCTV cameras. However, the surveillance infrastructure we were able to procure was part of a closed system making it impossible to connect CCTV output with parametric FRT within the limited resources of the project. A high-performance USB streaming camera[8] was then chosen. The camera connected to a MacBook laptop running Photo Booth to record videos in the .mov format to preserve the high quality of the camera video output. The videos were then trimmed to the intended clip, 2 frames per second were extracted ready to be processed by the FRT following the steps described in 3.1, namely detection, alignment, extraction and matching.

On arrival, participants were informed about the project and the experiment; it was made clear the recordings would not be shared with anyone and that commercial systems would not be used to guarantee data privacy. After signing the consent form each participant was first photographed front-face on a white background to create the database of identities; then they were shown the walking route and asked to move in a natural way while the camera-laptop setup recorded a video. In the first walk they did not wear any disguise (*disguise:clean* condition). They wore each disguise in turn and walked corridors and rooms toward the camera before returning to change disguise.

Participation in the evaluation was voluntary. White participants were recruited among colleagues from design, social science, and humanities aged 23 to 74. Black and brown participants were all international students attending master courses, were recruited in the university hall and received a £20 voucher as a thank you gift for their time (20-30 min)[9].

5.3 FRT performance evaluation

We carried out 2 evaluations, reported below. The first assessed a broader setup of 3 environments and 2 FRT systems; the findings were used to focus the larger second study on the system and the environment with the best performance while expanding the number of participants and their ethnicity.

5.3.1 Exploratory evaluation

The motivation underpinning the research was to study FRT in realistic conditions. We setup three recording stations that simulate realistic situations. The variable *environment: corridor, hall,* or *gate* depends on the position of the camera with respect to the person to be recognised (Fig. 12).

- The *hall* is a large room with a 4m ceiling; the camera was placed high, in the corner opposite the entrance door to get a wide-angle view of the whole space; the camera recorded participants entering the hall (from the corridor), walking toward the camera then turning around a large meeting table before leaving the room. The position of this camera matches CCTV monitoring public buildings, halls and the street.
- The *corridor* has a 2.5m ceiling and a series of spotlights that resulted in a rapidly changing illumination; the camera was places above a door frame pointing down to capture participants walking the corridor. This set-up resembles public passageways in cities, underground, or train stations.
- In the access control *gate* the camera is placed front face as in a passport control or private entrance.

[8] Technical details: full HD video 1080P, 60 FPS, 64 megapixels still resolution, wide-angle lens, auto-focus.

[9] Testing the disguises lasted about 10 minutes, the 20-30 minutes include a short interview that assessed participants' attitude towards FRT. The analysis of the interview is not included in this paper.

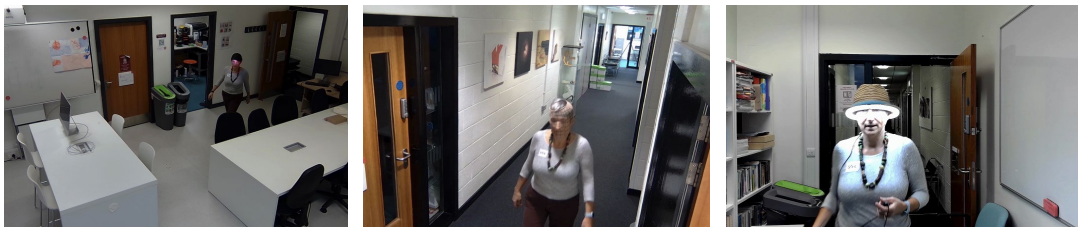


Fig. 12 The scenes recorded in the three settings in the exploratory evaluation: the hall, the corridor, and the access control gate.

Beside different camera-person settings, in the first evaluation we tested the open-source system DeepFace[10]. While the performance of open-source FRT may be below that of proprietary services such as Amazon Rekognition or Microsoft Azure, it guarantees the privacy of the participants' data. Deepface was used 'as is' with its pre-trained models Facenet512 and VGG-Face; we did not do any fine tuning to improve the systems performance on our data.

Fourteen (14) *white* participants took part in this evaluation; they walked the 3 *environments* with 11 *disguises*; the video clips were then computed. There is a notable discrepancy in the performance of both models when comparing the *environment:gate* with *environment:hall* and *environment:corridor*. In the gate scenario, both models demonstrate robust accuracy scoring 1 for the *clean* condition and achieving above 0.7 for seven distinct disguise types. This level of accuracy is significantly diminished in the other environments. For FaceNet-512, an accuracy above 0.7 is only observed for the *clean* and *halohat-off* in the *environment:hall*. Similarly, in the *environment:corridor,* this level of accuracy is reached only for the *clean, halohat-off,* and *snood_lines,* again specifically for FaceNet-512 (Fig. 13). The poor performance could be the combination of multiple factors: the corridor had sharp light-dark illumination due to a sequence of spots lights in the ceiling; in the hall the camera was placed high looking down with a skewed angle that may have obscured some facial features. It might be that a more powerful camera could have improved the performance or that, instead, for these challenging environments a different technique such as gait recognition would be more effective as it works with low-resolution video and does not need to detect the face [Wang et al. 2018].

In general, the disguises were effective in all environments although at different degrees. The good performance in *environment:gate* pushed us to focus on this setting as it puts the disguises in the most challenging conditions (i.e., the best conditions for FRT) giving us a better understanding of bypassing FRT.

5.3.2 Extended evaluation

In the second evaluation phase (July 2023) we engaged more participants of different ethnicities (*white*, *brown*, *black*). All the 11 *disguises* were used by 39 participants*: 22 white, 9 brown, 8 black* for a total of 429 video clips. The sample is unbalanced towards *white* (56%) participants respect to participants of colour (44%; *brown* 23%, *black* 21%), while it is balanced for gender (20 male and 19 female).

We used DeepFace and tested different parameters to generalise the results and avoid algorithm-specific limitations or idiosyncrasies. We selected three parameters for the two steps of (1) face detection and (2) face identification (see 3.1): (1) detector *backend*: *ssd, opencv, retinaface*; (2) face identification recognition *model*: *VGG-Face*, *Facenet*, *Facenet512*, *OpenFace*, *DeepFace*, and *ArcFace*; and (2) face identification distance *metrics*: *cosine*, *Euclidean*, and *euclidean_l2*. The combinations of the values of the 3 parameters (backend, algorithm, and metrics) generate 54 different configurations for the DeepFace system. All 429 video clips were tested with each configuration resulting in 23,166 single datapoints of performance (system configuration 54 with 11 disguises worn by 39 participant).

[10] DeepFace has been developed by Sefik Ilkin Serengil https://github.com/serengil/deepface

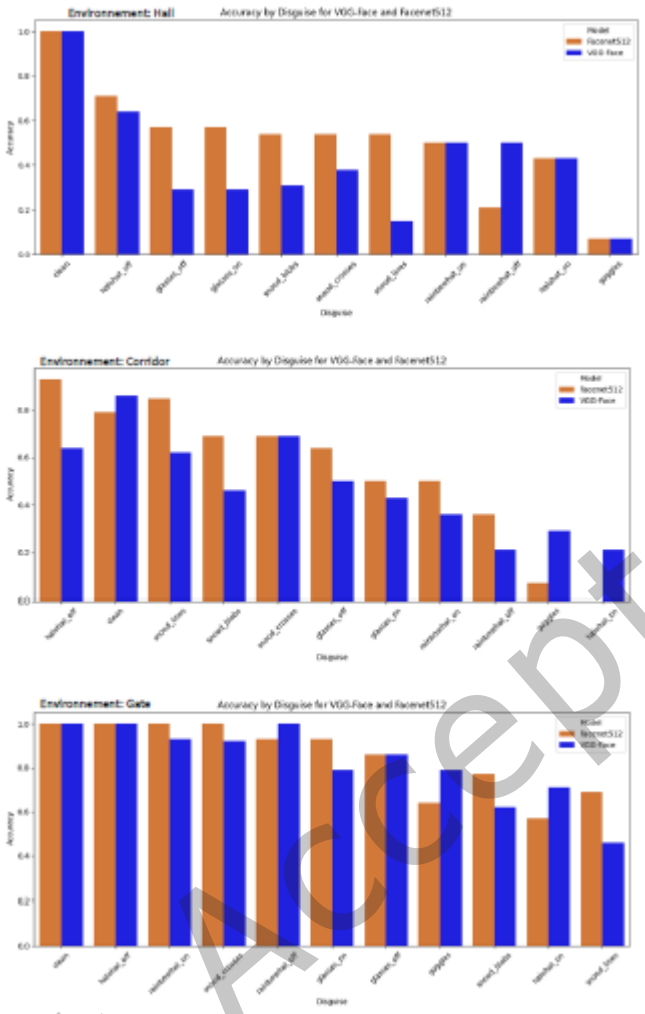


Fig. 13 Evaluation with 3 environments (hall, corridor, gate) and 11 disguises on DeepFace with Facenet512 and VGG-Face models.

The allocation of an identity to a video is done by majority votes: for each video, the occurrences of all the faces found in the frames are counted and the winner is decided by majority. For example, if the system detects 11 faces in the video and P1 is recognised 8 times while P2 is recognised 3 times, then P1 is selected as the person recognised in the video. In the tables, the columns Total and Count report the overall number of faces detected (total) and the number of times the person assigned to the video has been identified in the frames (count). The disguises can potentially have two effects:

- *to prevent face detection (evasion)*: if evasion occurs, no face is detected in a frame where, instead, there is a face; this is shown by the Total number of times a face is recognised: the lower the Total, the higher the evasion.
- *to increase the confusion of the system (increase the chances of dodging)*: the system is confused when more than one person is detected in a video, i.e., the closer Count and Total are, the less confused the system is. In the example above, for P1=8 and P2=3 the certainty is 8/11=0.72; if three people are identifies among 11 faces P3=5, P4=4, P5=2 then P3 is the winner and certainty decrease to 5/11=0.45 showing the disguise has been effective in confounding the FRT.

Less confusion does not equate to correct identification: the winning identity may or may not be the expected identity, i.e., the correct identity of the person in the video. The successful identification is given by 'accuracy', a standard measure used to evaluate machine learning systems defined as the number of correct predictions

divided by the number of predictions[11]. The tables below are ordered by decreasing 'accuracy'. To simplify the comparison and interpretation of the results, we report and discuss only the 10 best performing configurations. Extended tables of results can be found in Appendix A.1.

Table 1. White participants with no disguise (clean): best 10 performing configurations (extended table in Appendix A.1).

Model	Distance Metric	Backend	Total	Count	Ratio	Accuracy
Facenet512	euclidean_l2	retinaface	391	242	0.62	1.0
Facenet512	euclidean	retinaface	414	225	0.54	0.95
ArcFace	cosine	retinaface	367	185	0.50	0.77
ArcFace	euclidean_l2	retinaface	361	184	0.50	0.77
Facenet512	euclidean	opencv	465	202	0.43	0.77
Facenet512	euclidean_l2	opencv	448	221	0.49	0.77
ArcFace	euclidean	retinaface	374	172	0.45	0.77
Facenet	cosine	retinaface	329	164	0.41	0.73
Facenet512	cosine	retinaface	222	126	0.56	0.73
VGG-Face	cosine	retinaface	409	200	0.49	0.73

We first assess the performance on *disguise:clean*, the benchmark where participants did not wear any disguise, tabled by *ethnicity*. Unsurprisingly, the best performance is achieved for the *white*, although there was a wide variation from the first 2 configurations that achieve 100% and 95% and the 8-10 places with 73% (Table 1). With *brown* the best performance drops to 89% (Table 2), a result consistent across the top 23 positions occupied by the same 3 *models* (Facenet, Facenet512, VGG-face) albeit with different *metric* and *backend*. The top performance for *black* is 88% for the first 2 positions, comparable to that of the brown participants, but drops to 75% from the third position down (Table 3). The top 2 score across all ethnicities are between 88% and 100% thus in line with the performance reported in a 2022 study carried out using a FR commercial system (NEC Neoface V4 using HD5 Face Detector) to assess recordings done in the streets in the UK [Mansfield 2023]. Although there is a substantial difference in the scale of the database of identities (39 for us, 10,000 and 1,000 in [Mansfield 2023]) and the much more complex scene of a crowd walking along a street vs. a single individual walking toward a camera, the close result seems to indicate that the performance of open-source FRT and our experimental settings are a good approximation of a real situation and therefore a credible benchmark to check the effectiveness of the disguises.

[11] To compute the correctness of results we used Scikit-learn, an open-source Python library providing a wide range of tools and algorithms supporting a number of machine learning tasks and their evaluation.

Table 2. Brown participants with no disguise (clean): best 10 performing configurations (extended table in Appendix A.1).

Model	Distance Metric	Backend	Total	Count	Ratio	Accuracy
Facenet	cosine	opencv	186	121	0.65	0.89
Facenet512	cosine	ssd	172	116	0.67	0.89
Facenet	euclidean	opencv	194	120	0.61	0.89
VGG-Face	euclidean_l2	ssd	203	121	0.60	0.89
VGG-Face	cosine	ssd	204	121	0.59	0.89
VGG-Face	euclidean_l2	retinaface	202	146	0.72	0.89
VGG-Face	euclidean	retinaface	206	130	0.63	0.89
VGG-Face	cosine	retinaface	204	147	0.72	0.89
VGG-Face	euclidean_l2	opencv	208	134	0.64	0.89
VGG-Face	euclidean	opencv	209	116	0.55	0.89

Table 3. Black participants with no disguise (clean): best 10 performing configurations (extended table in Appendix A.1).

Model	Distance Metric	Backend	Total	Count	Ratio	Accuracy
Facenet512	euclidean	opencv	168	94	0.56	0.88
Facenet512	euclidean	retinaface	162	108	0.66	0.88
Facenet512	euclidean_l2	opencv	164	87	0.53	0.75
VGG-Face	euclidean_l2	retinaface	162	97	0.59	0.75
Facenet512	euclidean	ssd	160	87	0.54	0.75
Facenet512	euclidean_l2	retinaface	158	98	0.62	0.75
VGG-Face	cosine	retinaface	162	97	0.60	0.75
Facenet	cosine	opencv	152	72	0.47	0.75
ArcFace	cosine	ssd	156	77	0.49	0.75
ArcFace	euclidean_l2	retinaface	156	81	0.52	0.75

These three tables show that best models were more effective on white participants but more consistent with brown participants. The accuracy on black participants was lower than for white ones and more inconsistent than for brown ones. The most accurate configuration is *Facenet512, Euclidean, RetinaFace* occurring in the top 2 for both *white* and *black* and has the same top score of 0.89 accuracy for *brown*, see extended table in Appendix A.1.

Across all three ethnicities there are outliers, individuals who are easier or more difficult to recognise. For example, Figure 14 shows that, among the white, 015wtm (white male) is 8 times more difficult to identify than the easiest 014wtf (white female); among the brown 022brf (brown female) is 5 times more difficult than 037brf (brown female); and among the black 021blf (black female) is 7 times more difficult than 032blm (black male). Among all participants only 1 brown female was always correctly identified. Checking the video clips, we cannot find any obvious explanation due to the participants' behaviour. Indeed, one may expect that participants who did not pause much in front of the camera would be more difficult to recognise, yet all the three participants more difficult to identify paused for a few seconds before turning back giving the camera plenty of time to autofocus and to record good images. We may then conclude that some faces are more challenging to identify than others.

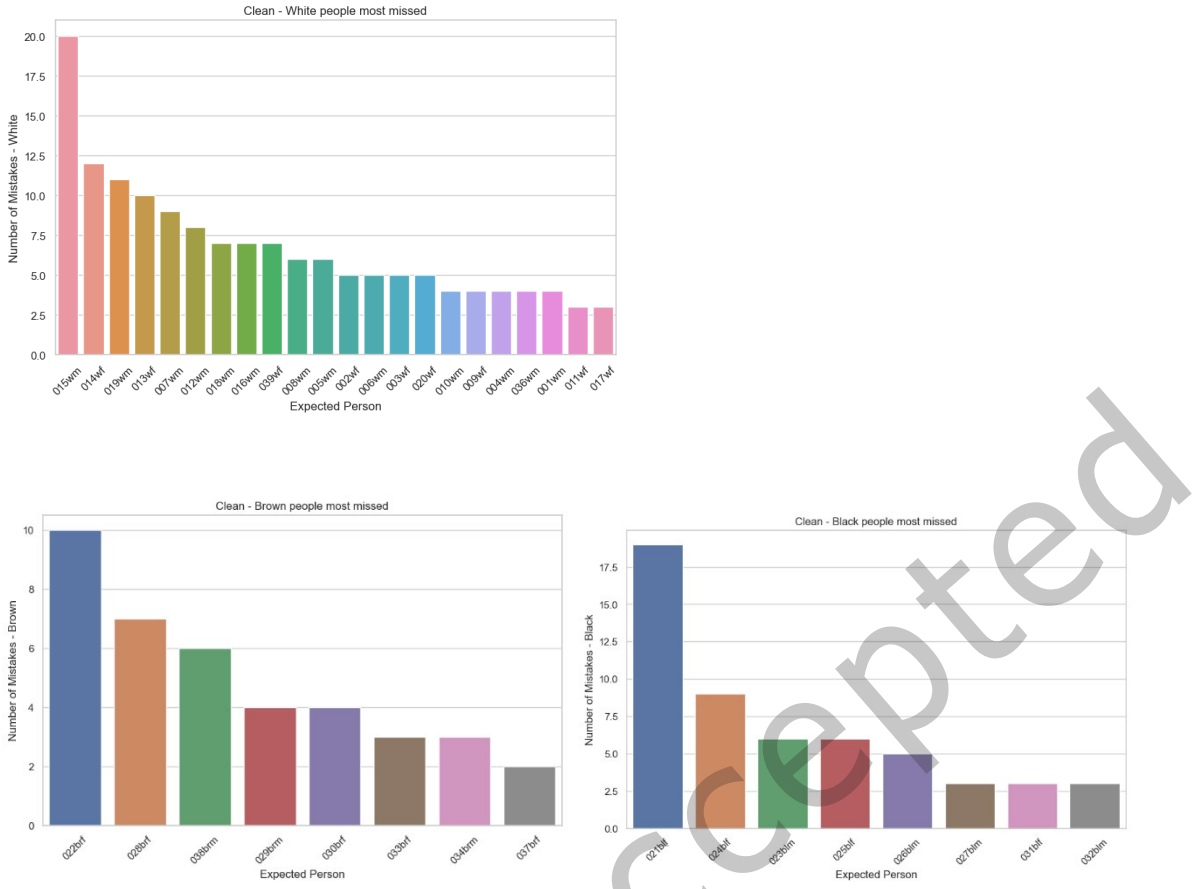


Fig. 14 The most difficult to the easiest person to identify among *white* (top), *brown* (left bottom), and *black* (right bottom) participants.

We then looked at who was most misrecognised, i.e. recognised as somebody else (Fig. 15). Unexpectedly the participant who is most frequently wrongly identified when a white person is in the video is a bearded brown man 034brm (55 times); the second is a brown female 037brf (20 times) followed by a white man 004wtm (17 time). A few more mistaken identities across ethnicities occur including white to black and vice versa. We could speculate why this is the case. In the process of facial features extraction, the coloured image is processed into levels of grey losing the skin colour and making brown skin closer in colour to South European or Middle Eastern people. The fact that a single individual has more than double the probability of being mistaken for others is significant in light of cases of mistaken identities.

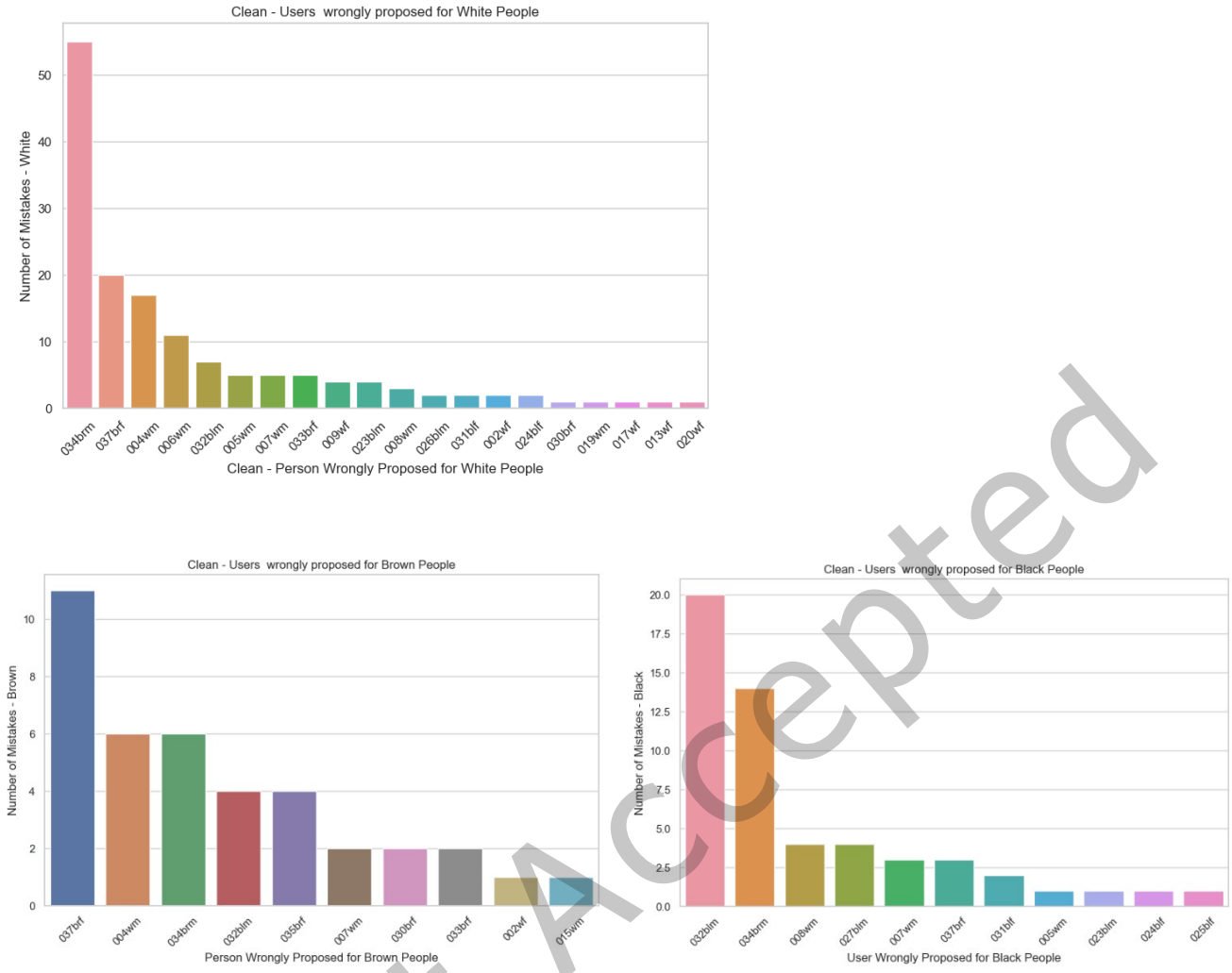


Fig. 15 The number of mistakes made in recognizing *white* (top), *brown* (bottom left), and *black* (bottom right).

Tables 4, 5, and 6 report the accuracy in correctly identifying the person when wearing a disguise. Comparing these tables with Tables 1, 2, and 3, the accuracy for *clean,* we see a sharp drop of about 20% in the performance with *white* participants from the very beginning while there is no difference for *brown* and *black* for the highest accuracy. Similarly, the *disguises:glass* | *halohat* | *rainbowhat* do not impact the performance for *brown* and *black* as much as for the *white*: the accuracy for *white* drops well below that of both *brown* and *black*.

The 4 best disguises across *all ethnicities* are the *snoods* and the *goggles* with a fall below 50% recognition. Comparing the ethnicities, it is evident that *halohat_on* is effective on the face of *white* while it has the opposite effect for *black* most likely because the bright light on the face makes facial features of black people more evident and therefore easier to be recognised. At the opposite, the bright light on white people makes features such as the nose and the face contour disappear thus effectively bypassing FRT.

Table 4. Effectiveness of the disguises for *white* participants. The most effective disguises in evading FRT are at the bottom.

Disguise	Model	Distance Metric	Backend	Total	Count	Ratio	Accuracy
clean	Facenet512	euclidean_l2	retinaface	391	242	0.62	1.0
halohat_off	Facenet512	euclidean_l2	retinaface	374	185	0.49	0.82
glasses_off	Facenet512	euclidean_l2	retinaface	316	139	0.43	0.73
rainbowhat_on	Facenet512	euclidean_l2	retinaface	389	161	0.41	0.68
rainbowhat_off	Facenet512	euclidean_l2	retinaface	328	136	0.41	0.64
glasses_on	ArcFace	cosine	retinaface	348	111	0.31	0.59
halohat_on	ArcFace	cosine	retinaface	397	123	0.30	0.55
snood_crosses	VGG-Face	cosine	retinaface	378	122	0.32	0.48
goggles	DeepFace	cosine	retinaface	80	25	0.31	0.4
snood_blobs	DeepFace	euclidean_l2	opencv	6	2	0.33	0.33
snood_lines	DeepFace	cosine	opencv	42	10	0.23	0.18

Table 5. Effectiveness of the disguises for brown participants. The most effective disguises in evading FRT are at the bottom.

Disguise	Model	Distance Metric	Backend	Total	Count	Ratio	Accuracy
clean	Facenet	cosine	opencv	186	121	0.65	0.89
glasses_off	Facenet	cosine	retinaface	170	105	0.61	0.89
glasses_on	Facenet	euclidean	retinaface	196	133	0.68	0.89
halohat_off	Facenet512	euclidean_l2	opencv	196	111	0.56	0.89
rainbowhat_on	ArcFace	cosine	opencv	186	109	0.59	0.89
halohat_on	ArcFace	cosine	retinaface	221	120	0.54	0.78
rainbowhat_on	ArcFace	cosine	retinaface	192	119	0.61	0.78
snood_crosses	Facenet	cosine	retinaface	232	90	0.39	0.56
goggles	DeepFace	cosine	retinaface	98	14	0.14	0.44
snood_blobs	ArcFace	cosine	opencv	188	44	0.23	0.33
snood_lines	DeepFace	euclidean	opencv	318	168	0.52	0.33

Table 6. Effectiveness of the disguises for black participants. The most effective disguises in evading FRT are at the bottom.

Disguise	Model	Distance Metric	Backend	Total	Count	Ratio	Accuracy
clean	Facenet512	euclidean	opencv	168	94	0.56	0.88
halohat_off	ArcFace	cosine	retinaface	152	95	0.62	0.88
halohat_on	ArcFace	cosine	retinaface	186	106	0.57	0.75
rainbowhat_on	ArcFace	cosine	retinaface	166	119	0.71	0.75
glasses_off	ArcFace	cosine	retinaface	152	94	0.62	0.62
glasses_on	Facenet512	euclidean	retinaface	201	76	0.38	0.62
rainbowhat_off	ArcFace	cosine	retinaface	157	102	0.65	0.62
snood_crosses	ArcFace	cosine	retinaface	178	94	0.52	0.50
goggles	Facenet512	euclidean_l2	retinaface	181	70	0.39	0.43
snood_blobs	ArcFace	euclidean	retinaface	176	34	0.19	0.25
snood_lines	DeepFace	euclidean	opencv	86	14	0.16	0.25

In tables 4-6 we present the best performing settings for each disguise. This is unrealistic as in a real-world scenario where only one configuration is used most likely reducing accuracy across the different disguises. For example, quite a few configurations report zero accuracy when confronted with the snood disguises.

Looking into the confounding effect of the disguises some interesting phenomena emerge. For *white* in *clean*, 015wtm in the most misrecognised; in *disguises* 015wtm is joined by 10wtm and 19wtm as people most difficult to identify when wearing *disguises*. Moreover, the number of mistakes increases dramatically across the whole sample (Fig. 16). *Brown* and *black* show a similar trend when comparing *clean* and *disguises* (see appendix 2.A).

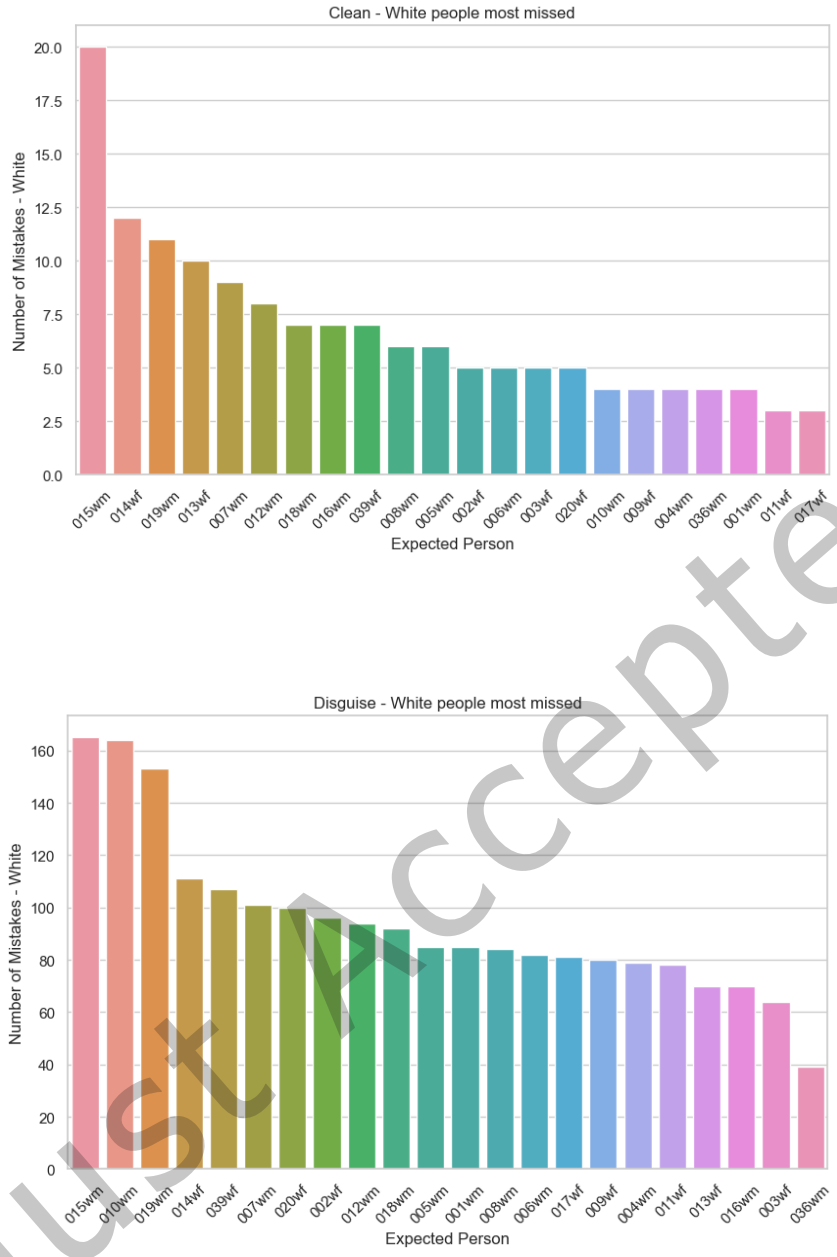


Fig. 16 The number of mistakes made in recognizing white in the benchmark clean (top) and across all the disguises (bottom).

Even if not every disguise was effective in the same way across *ethnicities* (see Tables 4, 5, 6), taken together they were effective in confounding FRT. Fig. 17 shows that the disguises confound ethnicity for *brown* with all three ethnicities (white, brown, and black) all scoring as the three most frequent mistakes. This phenomenon is less marked for *black* even if some *white* are mistaken for black (Fig. 17 bottom). Fig. 17 top shows participant 034brm, the bearded brown man wrongly identified most often as white in the *clean* (55 times, Fig. 15) shoots to about 950 mistakes when *disguises* are worn making the case of mistaken identities (due to dodging) a serious issue. The most reasonable explanation is that somehow 034brm has a 'common face'.

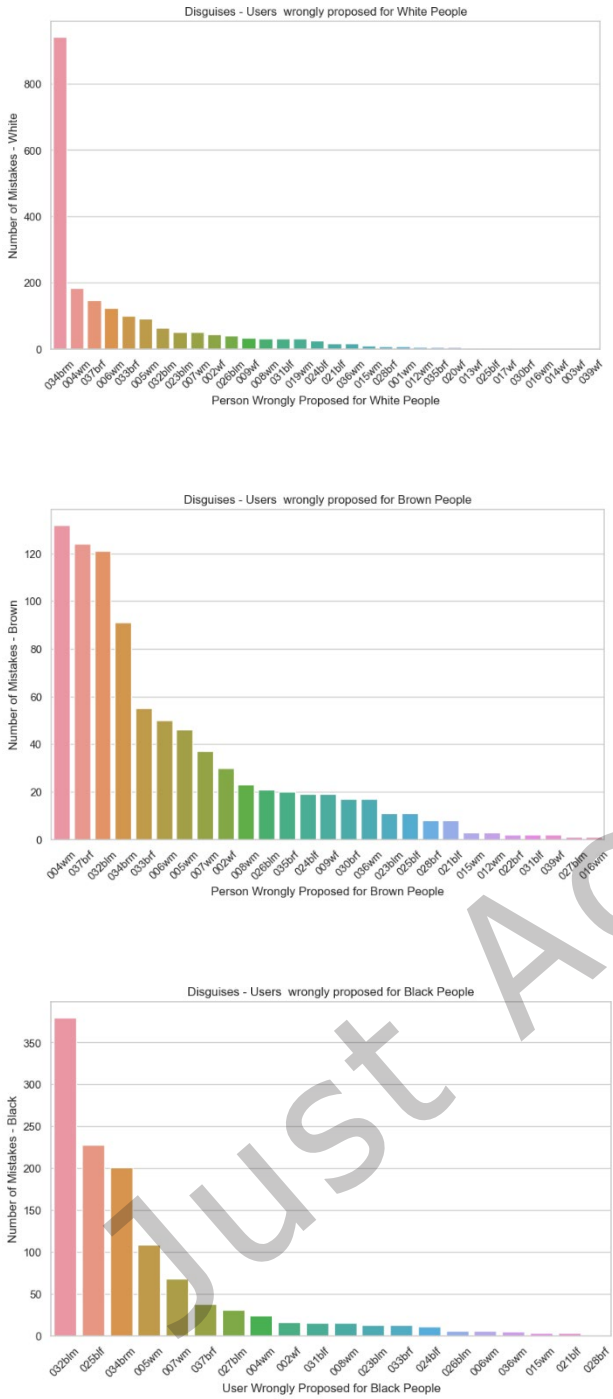


Fig. 17 The number of mistakes made in recognizing white (top), brown (centre), and black (bottom) when the disguises are worn.

The creation of a 3D model of the face is an essential part of the FRT process [Kortli et al. 2020]. To complete our empirical exploration, we created a second database of identities adding to the one created out of 1 high-quality photo per participants, a further 4 images of the person in different positions extracted from the 'clean' video. The rationale was that images with different face positions would improve the recognition performance. For this test only the two best performing configurations were used, namely *Facenet512-euclidean-retinaface*

and *Facenet512-euclidean_l2-retinaface*, on 39 participants with *environment:gate*. Unexpectedly, the comparison points in the opposite direction with a 20% drop in accuracy across the three ethnicities in the *clean* benchmark. We hypothesize that this is because the increased number of images and positions of the face reduces somehow the distance among the photos of different participants in the database making the process of recognition harder. Should this be the case, this could point towards a problem for systems using very large databases. There is only a limited loss, if any, in the *disguises* condition, most likely because the disguises had already drastically reduced the performance. The tables are in Appendix A.3.

5.3.3 Logistic regression analysis

A Likelihood Ratio Test (LRT) was used to evaluate the impact of each variable on the predictive capacity of one of the top-performing FRT pipelines, specifically *Facenet512-retinaface-euclidean_l2*. This was achieved by comparing two logistic regression models. The full model consists of a logistic regression in the form: 'Result ~ Gender + Ethnicity + Disguise'. This is compared with three reduced models where in each one of the regressors is removed. The p-value from these tests indicates whether that variable is useful in explaining the predictive performance of the FRT pipeline. We formulate the following hypothesis:

- Null hypothesis H_0: if the full model and the nested model fit the data equally then the removed variable is not significant to the capacity of the FRT model to recognize or not a person.
- Alternative hypothesis H_A: if the full model fits the data better than the nested model, so the variable is significant to the capacity of the FRT model to recognize or not a person.

The p-values from our LRT analyses are as follows:

- "*Result ~ Gender + Ethnicity*" (reduced model): $1.58\ x\ 10^{-173}$
- "*Result ~ Ethnicity + Disguise*" (reduced model): $3.78\ x\ 10^{-70}$
- "*Result ~ Gender + Disguise*" (reduced model): $1.31\ x\ 10^{-8}$

The LRT test for all three models yields p-values less than 0.05. This leads us to reject the null hypothesis, pointing to *Disguise*, *Ethnicity* and *Gender* as significant contributors to the FRT pipeline's capacity to predict.

Table 7 The coefficients (coef), p-value (P>|z|), and [0.025, 0.977] show 95% confidence interval (CI) of the logistic regression model, with the coefficients for *disguise* being compared to *clean* as reference level (one-hot encoding), the one for *gender* is compared to *female* and *ethnicity* compared to *black*.

Feature	Coefficient	P>|z|	[0.025	0.975]
Intercept	0.59	0.00	0.41	0.77
Gender[Male]	0.29	0.00	0.19	0.40
Ethnicity[Brown]	0.57	0.00	0.43	0.70
Ethnicity[White]	-0.53	0.00	-0.65	-0.40
Disguise[glasses_off]	-0.63	0.00	-0.84	-0.41
Disguise[glasses_on]	-0.76	0.00	-0.97	-0.54
Disguise[goggles]	-1.42	0.00	-1.65	-1.20
Disguise[halohat_off]	-0.24	0.26	-0.46	-0.03
Disguise[halohat_on]	-0.86	0.00	-1.07	-0.65
Disguise[rainbowhat_off]	-0.60	0.00	-0.82	-0.39
Disguise[rainbowhat_on]	-0.50	0.00	-0.71	-0.29
Disguise[snood_blobs]	-2.31	0.00	-2.60	-2.02
Disguise[snood_crosses]	-1.50	0.00	-1.73	-1.27
Disguise[snood_lines]	-3.00	0.00	-3.33	-2.66

We also include an analysis of the coefficients of the full model to assess how each level of the categorical variables influences the predictive accuracy of the FRT pipeline. For the three categorical variables we use as reference category *clean* for disguise, *female* for gender and *black* for ethnicity. The coefficients for the other levels of these variables represent the effect of that level on the predictive accuracy of the model compared to the reference level. The results in Table 7 point to a significant difference between the disguises compared to the reference and for the ethnicities.

The coefficients associated with the various disguises are negative. This means that the presence of disguises negatively impacts the FRT pipeline's predictive accuracy compared to *clean* faces. Notably, *snood_lines*, *snood_blobs*, *snood_crosses* and *goggles* exhibit the larges absolute impacts on prediction capacity. Ethnicity also showed a significant impact: compared to the *black* reference group, *white* ethnicity had a negative impact on performance, while *brown* ethnicity showed a positive impact. Finally, the gender *male* has a slightly positive impact on performance compared to the *female* reference group.

5.4 Testing FRT advancements

1.1.1 5.4.1 Comparison of models performance one year on

The extended evaluation reported in 5.3.2 used FRT available in December 2022. As facial recognition is improving fast in light of the evolution of AI, in January 2024 we run new tests to check if and how the performance of FRT had changed with the latest development. A year on, the models used in the extended evaluation had not changed while new models have been proposed. In selecting the models for the comparative test, we used variety as key factor and chose three new models to test:

- *QMagFace* [Terhörst et al. 2023] focuses on FRT in difficult conditions, combining a quality-aware comparison function with a face recognition model trained (MagFace) [Terhörst et al. 2023]. It assesses the quality of a face image (including factors such as lighting and pose) to enhance recognition accuracy. In training this helps the model to weight facial features based on their importance, improving performance, especially in challenging conditions like varying angles and image quality.
- *ElasticFace* [Boutros et al. 2022] uses a FR strategy based on the use of flexible margin loss which enhances the model's ability to closely group similar faces (intra-class compactness) while better distinguishing between different faces (inter-class discrepancy), helping the model to differentiate more effectively between face classes [Boutros et al. 2023].
- *GhostFaceNet* [Alansari et al. 2023] uses Ghost modules for generating feature maps with fewer parameters and less computational complexity. The architecture employs a modified Global Depthwise Convolution (GDC) for improved face feature representation. These advancements enable GhostFaceNets to achieve a balance between computational efficiency and high accuracy in facial recognition tasks [Alansari et al. 2023].

The three *models* were downloaded from their respective Git Hub repositories[12]. For *QMagFace*, *MTCNN* was used as the *backend* framework, as suggested by the developers. For *ElasticFace* and *GhostFaceNet* models, a combination of *RetinaFace*, *OpenCV*, and *ssd* were selected as the backend framework. These models were selected for analysis based on two key criteria: their superior performance across a wide array of public evaluation datasets and their distinction as significant advancements of the state-of-the-art FRT. Consequently, our primary aim with this round of tests is to evaluate whether these new models constitute a marked improvements over the models available in January 2023 on the same setup, datasets and videos described in section 5.3.2. The extended results are in Appendix B.

[12] QMagFace has been developed by pterhoer and mihlefeld: https://github.com/pterhoer/QMagFace

ElasticFace has been developed by Fadi Boutros : https://github.com/fdbtrs/ElasticFace

GhostFaceNet has been developed by HamadYA: https://github.com/HamadYA/GhostFaceNets

Tables 8, 9, and 10 compare the best 10 performances for *clean* (participants wearing no disguise) for the three ethnicities (*white:clean*; *brown:clean*; *black:clean*). As in 2023 (see 5.3.2), we calculate the total number of faces detected (Tot) and then we count how many times the person in the video was correctly identified (Corr). The closer T is to C, the more accurate (Acc) the system was in detecting the face and in recognising the person. Tables 8, 9 and 10 show a clear and marked improvement, with accuracy increasing for all three ethnicities. As a general observation, the performance of *QMagFace:Mtcnn* and *ElasticFace:retinaface* tend to pair, while the lightweight *GhostFaceNet* system does not perform well: *GhostFaceNet* appears only 1 time in the 10th position for *white:clean*, but does not make the grade for *brown:clean* and *black:clean* (see Appendix B for the results in full).

Table 8. The best 10 performing configurations in 2023 and 2024 for white ethnicity with no disguise (*white:clean*).

Evaluation 2023 *white:clean*				
System configuration	**T**	**C**	**T-C**	**A**
Facenet512 : euclidean_l2 :	391	242	149	1.0
Facenet512 : euclidean : retinaface	414	225	189	0.95
ArcFace : cosine : retinaface	367	185	182	0.77
ArcFace : euclidean_l2 : retinaface	361	184	177	0.77
Facenet512 : euclidean : opencv	465	202	263	0.77
Facenet512 : euclidean_l2 : opencv	448	221	227	0.77
ArcFace : euclidean : retinaface	374	172	202	0.77
Facenet : cosine : retinaface	329	164	165	0.73
Facenet512 : cosine : retinaface	222	126	96	0.73
VGG-Face : cosine : retinaface	409	200	209	0.73

Evaluation 2024 *white:clean*				
System configuration	**T**	**C**	**T-C**	**A**
QMagFace : euclidean : Mtcnn	394	282	112	1.0
QMagFace : euclidean_12 : Mtcnn	394	284	110	1.0
ElasticFace-Arc+ : cosine : retinaface	403	245	158	0.95
ElasticFace-Arc+ : euclidean_12 :	403	245	158	0.95
ElasticFace-Arc+ : cosine : opencv	403	239	164	0.91
ElasticFace-Arc+ : euclidean_12 : opencv	403	239	164	0.91
QMagFace : cosine : Mtcnn	394	270	124	0.91
ElasticFace-Cos : cosine : retinaface	403	223	180	0.86
ElasticFace-Cos : euclidean_12 : retinaface	403	223	180	0.86
GhostFaceNet : euclidean_12 : retinaface	403	201	202	0.86

Table 9. The best 10 performing configurations in 2023 and 2024 for brown ethnicity with no disguise (*brown:clean*).

Evaluation 2023 *brown:clean*				
System configuration	**T**	**C**	**T-C**	**A**
Facenet : cosine : opencv	186	121	65	0.89
Facenet512 : cosine : ssd	172	116	56	0.89
Facenet : Euclidean : opencv	194	120	74	0.89
VGG-Face : euclidean_l2 : ssd	203	121	82	0.89
VGG-Face : cosine : ssd	204	121	83	0.89
VGG-Face : euclidean_l2 :	202	146	56	0.89
VGG-Face : euclidean : retinaface	206	130	76	0.89
VGG-Face : cosine : retinaface	204	147	57	0.89
VGG-Face : euclidean_l2 : opencv	208	134	74	0.89
VGG-Face : uclidean : opencv	209	116	93	0.89

Evaluation 2024 *brown:clean*				
System configuration	**T**	**C**	**T-C**	**A**
ElasticFace-Arc+ : cosine : retinaface	205	145	60	1.0
QMagFace : cosine : Mtcnn	200	170	30	1.0
QMagFace : euclidean : Mtcnn	200	174	26	1.0
QMagFace : eucliidean_12 : Mtcnn	200	177	23	1.0
ElasticFace-Arc+ : euclidean_12	205	145	60	1.0
ElasticFace-Cos : euclidean_12 : opencv	205	134	71	0.89
ElasticFace-Cos : euclidean_12 :	205	137	68	0.89
ElasticFace-Cos : cosine : retinaface	205	137	68	0.89
ElasticFace-Cos : cosine : opencv	205	134	71	0.89
ElasticFace-Arc+ : euclidean_12 : openvc	205	142	63	0.89

Table 10. The best 10 performing configurations in 2023 and 2024 for black ethnicity with no disguise (*black:clean*).

Evaluation 2023 *black:clean*				
System configuration	**T**	**C**	**T-C**	**A**
Facenet512 : Euclidean : opencv	168	94	74	0.88
Facenet512 : euclidean : retinaface	162	108	54	0.88
Facenet512 : euclidean_l2 : opencv	164	87	77	0.75
VGG-Face : euclidean_l2 : retinaface	162	97	65	0.75
Facenet512 : euclidean : ssd	160	87	73	0.75
Facenet512 : euclidean_l2 :	158	98	60	0.75
VGG-Face : cosine : retinaface	162	97	65	0.75
Facenet : cosine : opencv	152	72	80	0.75
ArcFace : cosine : ssd	156	77	79	0.75
ArcFace : euclidean_l2 : retinaface	156	81	75	0.75

Evaluation 2024 *black:clean*				
System configuration	**T**	**C**	**T-C**	**A**
ElasticFace-Cos : cosine : retinaface	162	111	51	1.0
ElasticFace-Cos: euclidean_l2: retinaface	162	111	51	1.0
ElasticFace-Arc+ : cosine: opencv	162	93	69	0.88
ElasticFace-Arc+ : euclidean_l2 : retinaface	162	93	69	0.88
ElasticFace-Arc+ : cosine : retinaface	162	97	65	0.88
ElasticFace-Arc+ : euclidean_12 :	162	975	65	0.88
ElasticFace-Cos : cosine : opencv	162	105	57	0.88
ElasticFace-Cos : euclidean_l2 : opencv	162	105	57	0.88
QMagFace : cosine : Mtcnn	159	121	38	0.88
QMagFace : euclidean : Mtcnn	159	124	35	0.88

There has been a sharp improvement in performance with *clean* for *brown* and *black*. There is also an increase, albeit to a lower degree when participants wear the disguises (Tables 11, 12 and 13). However, there is a considerable difference in the system rankings with *QMagFace* firmly in the lead. Remarkably this happens with several 100% results for *brown* and *black* participants which, in the past, have been difficult faces to recognise. The reason for *QMagFace* outperforming the other systems may be the intentional goal to cope with difficult images, possibly rotated heads and partially obstructed faces. The result seems to be a system able to infer an identity from very limited visible facial features.

An unexpected result is the higher performance with *brown* and *black* than with *white* participants. This could be explained by the sample size as the white participants (22) are nearly 3 times the brown (9) and black (8). A small sample set means each may be easily discriminated by the neural networks so that they look distant from the others in the search space and hence they are more easily discriminated. As the sample size increases even slightly (as in the case of the *white*), the distance between the individuals decreases, therefore potentially inducing more mistakes when the face is partially concealed. Our interpretation is then that, while in 2023 the disguises were disruptive for the performance of the systems that often missed to detect the face entirely, in 2024 the systems are able to detect and discriminate faces form just a few facial features when the sample is small as there is enough information to separate the individuals in the image reference set. However, as the sample increases those few features are not enough to discriminate between individuals and the performance decreases. We could then expect that the same reduction in performance with *white* would occur with a larger sample of *brown* and *black* participants.

Table 11. Comparison of the performance of the systems against the disguises in 2023 and 2024 for white ethnicity.

Evaluation 2023 *white : disguises*						Evaluation 2024 *white : disguises*				
Rank	**System configuration**	**T**	**C**	**A**	**Disguise**	**Rank**	**System configuration**	**T**	**C**	**A**
0	Facenet512:euclidean_12:retinaface	391	2	1.0	clean	0	QMagFace : euclidean : Mtcnn	394	282	1.0
1	Facenet512:euclidean_12:retinaface	347	1	0.82	halohat_off	3	QMagFace : euclidean_l2 : Mtcnn	388	243	0.91
2	Facenet512:euclidean_12:retinaface	316	1	0.73	glasses_off	5	ElasticFaceArc+:cosine:retinaface	375	167	0.86
3	Facenet512:euclidean_l2:retinaface	389	1	0.68	rainbowhat_on	6	QMagFace : euclidean_l2 : Mtcnn	424	203	0.76
4	Facenet512:euclidean_l2:retinaface	328	1	0.64	rainbowhat_off	1	QMagFace:euclidean_l2:Mtcnn	339	205	1.0
5	ArcFace : cosine : retinaface	348	1	0.59	glasses_on	8	ElasticFaceArc+:cosine:retinaface	365	150	0.73
6	ArcFace : cosine : retinaface	397	1	0.55	halohat_on	10	QMagFace:euclidean_l2:Mtcnn	428	142	0.45
7	VGG-Face : cosine : retinaface	378	1	0.48	snood_crosses	2	QMagFace:euclidean:Mtcnn	364	216	0.95
8	DeepFace : cosine : retinaface	80	25	0.4	goggles	9	ElasticFaceArc+:cosine:retinaface	370	128	0.68
9	DeepFace : euclidean_12 : opencv	6	2	0.33	snood_blobs	4	QMagFace:euclidean_l2:Mtcnn	188	351	0.9
10	DeepFace : cosine : opencv	42	10	0.18	snood_lines	7	QMagFace : euclidean_l2 : Mtcnn	361	173	0.76

Table 12. Comparison of the performance of the systems against the disguises in 2023 and 2024 for brown ethnicity.

Evaluation 2023 *brown : disguises*						Evaluation 2024 *brown : disguises*				
Rank	**System configuration**	**T**	**C**	**A**	**Disguise**	**Rank**	**System configuration**	**T**	**C**	**A**
0	Facenet : cosine : opencv	186	121	0.89	clean	0	QMagFace : cosine : Mtcnn	200	170	1.0
1	Facenet : cosine : retinaface	170	105	0.89	glasses_off	5	QMagFace : cosine : Mtcnn	196	164	1.0
2	Facenet : euclidean : retinaface	196	133	0.89	glasses_on	8	QMagFace : cosine : Mtcnn	211	166	1.0
3	Facenet512 : euclidean_l2 : opencv	196	111	0.89	halohat_off	3	QMagFace : cosine : Mtcnn	186	145	1.0
4	Arcface : cosine : opencv	186	109	0.89	rainbowhat_on	7	QMagFace : euclidean : Mtcnn	235	187	1.0
5	ArcFace : cosine : retinaface	221	120	0.78	halohat_on	4	QMagFace : euclidean_12 : Mtcnn	201	142	1.0
6	ArcFace : cosine : retinaface	192	110	0.78	rainbowhat_off	6	QMagFace : cosine : Mtcnn	181	143	1.0
7	Facenet : cosine : retinaface	232	90	0.56	snood_crosses	2	QMagFace : euclidean : Mtcnn	201	148	1.0
8	DeepFace : cosine : retinaface	98	14	0.44	goggles	9	QMagFace : euclidean_12 : Mtcnn	228	159	0.89
9	Arcface : cosine : opencv	188	44	0.33	snood_blobs	10	QMagFace : euclidean_12 : Mtcnn	197	154	0.89
10	DeepFace : euclidean : opencv	318	168	0.33	snood_lines	1	QMagFace : euclidean_12 : Mtcnn	215	163	1.0

Table 13. Comparison of the performance of the systems against the disguises in 2023 and 2024 for black ethnicity.

Evaluation 2023 *black : disguises*						Evaluation 2024 *black : disguises*				
Rank	**System configuration**	**T**	**C**	**A**	**Disguise**	**Rank**	**System configuration**	**T**	**C**	**A**
0	Facenet512 : Euclidean : opencv	168	9	0.88	clean	0	QMagFace : euclidean_12 : Mtccn	162	111	1.0
1	ArcFace : cosine : retinaface	152		0.88	halohat_off	2	QMagFace: euclidean_12 : Mtcnn	176	138	1.0
2	ArcFace : cosine : retinaface	186		075	halohat_on	5	QMagFace: euclidean_12 : Mtcnn	200	148	0.88
3	ArcFace : cosine : retinaface	166		0.75	rainbowhat_on	7	QMagFace: euclidean_12 : Mtcnn	191	141	0.88
4	ArcFace : cosine : retinaface	152		0.62	glasses_off	1	QMagFace: euclidean_12 : Mtcnn	197	155	1.0
5	Facenet512 : euclid : retinaface	201		0.62	glasses_on	4	ElasticFace-Cos : cosine : ssd	211	125	0.88
6	ArcFace : cosine : retinaface	157		0.62	rainbowhat_off	6	QMagFace: euclidean_12 : Mtcnn	172	130	0.88
7	ArcFace : cosine : retinaface	178		0.50	snood_crosses	3	QMagFace: euclidean_12 : Mtcnn	186	140	1.0
8	Facenet512:eucl_12:retinaface	181		0.43	goggles	10	ElasticFace-Arc+:cosine:	200	60	0.38
9	ArcFace : euclidean : retinaface	176		0.25	snood_blobs	9	ElasticFace-Arc+:cosine:	179	87	0.5
10	DeepFace : euclidean : opencv	86	14	0.25	snood_lines	8	ElasticFace-Arc+:cosine:	200	94	0.62

Fig. 18 shows the increase in performance from 2023 when *white* individuals were mostly missed to be recognised: for all ethnicities the number of missed identifications has consistently decreased meaning that the new systems make less mistakes. The most pronounced improvement is for *white* with the highest number nearly halved and the other reduced of about 20% (the graphs for *brown* and *black* are in Appendix B.2).

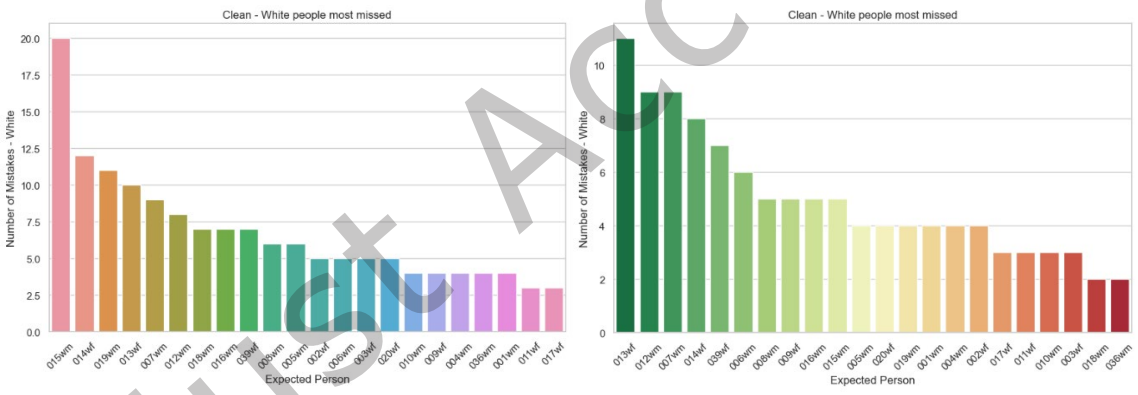


Fig. 18 A comparison of the *white* participants most missed to be identified: 2023 on the left, 2024 on the right.

The last comparison between the 2023 and 2024 evaluations is the misclassification of individuals wearing disguises. Fig. 19 compares the misclassification for the three ethnicities in 2023 and 2024. While for *white* and *black* there has been a decrease in misrecognitions, the values are still high. The people misrecognised have changed, but there is a consistence within the experimental setup meaning that the same small group of people is the most misrecognised within the evaluation of 2023 and another one is misrecognised in the evaluation of 2024. In both years misrecognition occurs across ethnicities in contradiction with the results shown in [Rosenberg et al. 2023] "that it is easier to impersonate identities in the same demographic" (pg.7236): our physical disguises were most effective across ethnicities in *white* and *brown* in 2023 and across *white* and *black* in 2024. The most misrecognised participants were 034brm in 2023 and 033brf both *brown* participants from India, a male and a female respectively. Several could be the reasons for this contradictory results. The individuals mostly misrecognised are both from India, a group poorly represented in [Rosenberg et al. 2023] that distinguish White (10.000 individuals), Asian (2.500), Black (1240) and Indian (20). Assuming 'Asian' includes ethnicities from China, Japan, and Korea, we could speculate that the 'impersonation within the same

demographic' is due to distinct facial traits for each demographic group White, Asian and Black and therefore easier to cluster, while people from India tend to have Caucasian facial features similar to those of white people but different in skin tone. Another potential explanation could be the process of generating computational perturbations from the same photos used for recognition [Rosenberg et al. 2023] that is radically different from our physical approach: most likely our disguises conceal or change the facial features in a radically different way and, therefore, our results are simply not comparable with computational perturbations. Indeed, [Rosenberg et al. 2023] pg. 7236 states that "With the exception of the Indian demographic group, which is too small to source any significant conclusions, minority groups have lower matching performance. This is attributed to the tightly clustered embeddings we observe for minority demographic groups." A further explanation could be the source material, live videos for us vs. photos for [Rosenberg et al. 2023], or because of different models and different processing pipeline. Indeed [Rosenberg et al. 2023] observe that demographics are discernible in the embedding phase with skin tones differentiated in early layers of the network. While we cannot say with certainty what the cause is, if, for example a single setting is responsible for the majority of misidentifications or if instead is one among the algorithms, the metrics, or the backends, the logistic regression provides some hindsight.

We conducted a Likelihood Ratio Test (LRT) to evaluate the impact of each variable on the predictive capacity of one of the recent top-performing FRT pipelines, specifically *QMagFace:mtcnn:euclidean_l2*. This was achieved by comparing two logistic regression models. The full model consists of a logistic regression of the form: '*Result ~ Gender + Ethnicity + Disguise'*. This is compared with three reduced models where in each one of the regressors is removed. The p-value from these tests indicates whether that variable is useful in explaining the predictive performance of the FRT pipeline. Our hypothesis are:

- Null hypothesis H_0: if the full model and the nested model fit data equally then the removed variable is not significant to the capacity of the FRT model to recognize or not a person.
- Alternative hypothesis H_A: if the full model fits data better then the nested model, so the variable is significant to the capacity of the FRT model to recognize or not a person.

The p-values from our LRT analyses are as follows:

- *'Result ~ Gender + Ethnicity'* (reduced model): $3.88\ x\ 10^{-73}$
- *'Result ~ Ethnicity + Disguise'* (reduced model): 0.44
- *'Result ~ Gender + Disguise'* (reduced model): $1.19\ x\ 10^{-99}$

The LRT test for the models of '*Result ~ Gender + Ethnicity*' and '*Result ~ Gender + Disguise*' both yield p-values less than 0.05. This leads us to reject the null hypothesis, pointing to *disguise* and *ethnicity* as significant contributors to the FRT pipeline's capacity to predict. Conversely, the '*Result ~ Ethnicity + Disguise*' LRT results in a high p-value (0.44), leading us to not reject the null hypothesis. This supports the conclusion that *gender* is not a significant factor in the recent FRT model's ability to recognize individuals in this context.

We also include an analysis of the coefficients of the full model to assess how each level of the categorical variables influences the predictive accuracy of the FRT pipeline. For the three categorical variables we use as reference category *clean* for *disguise*, *female* for *gender* and *black* for *ethnicity*. The coefficients for the other levels of these variables represent the effect of that level on the predictive accuracy of the model compared to the reference level. The results in Table 14 also point to significant differences between *disguises* compared to the reference and for *ethnicities*, while there is no significant difference in the predictive accuracy of the model for *male* compared to *female*. Significance is assessed at a 5% level of significance. The coefficients associated with the various *disguises* are negative meaning that the presence of *disguises* negatively impacts the FRT pipeline's predictive accuracy compared to *clean* faces. Notably, *googles*, *halohat_on*, and *snood_lines* exhibit the largest absolute impacts on prediction capacity.

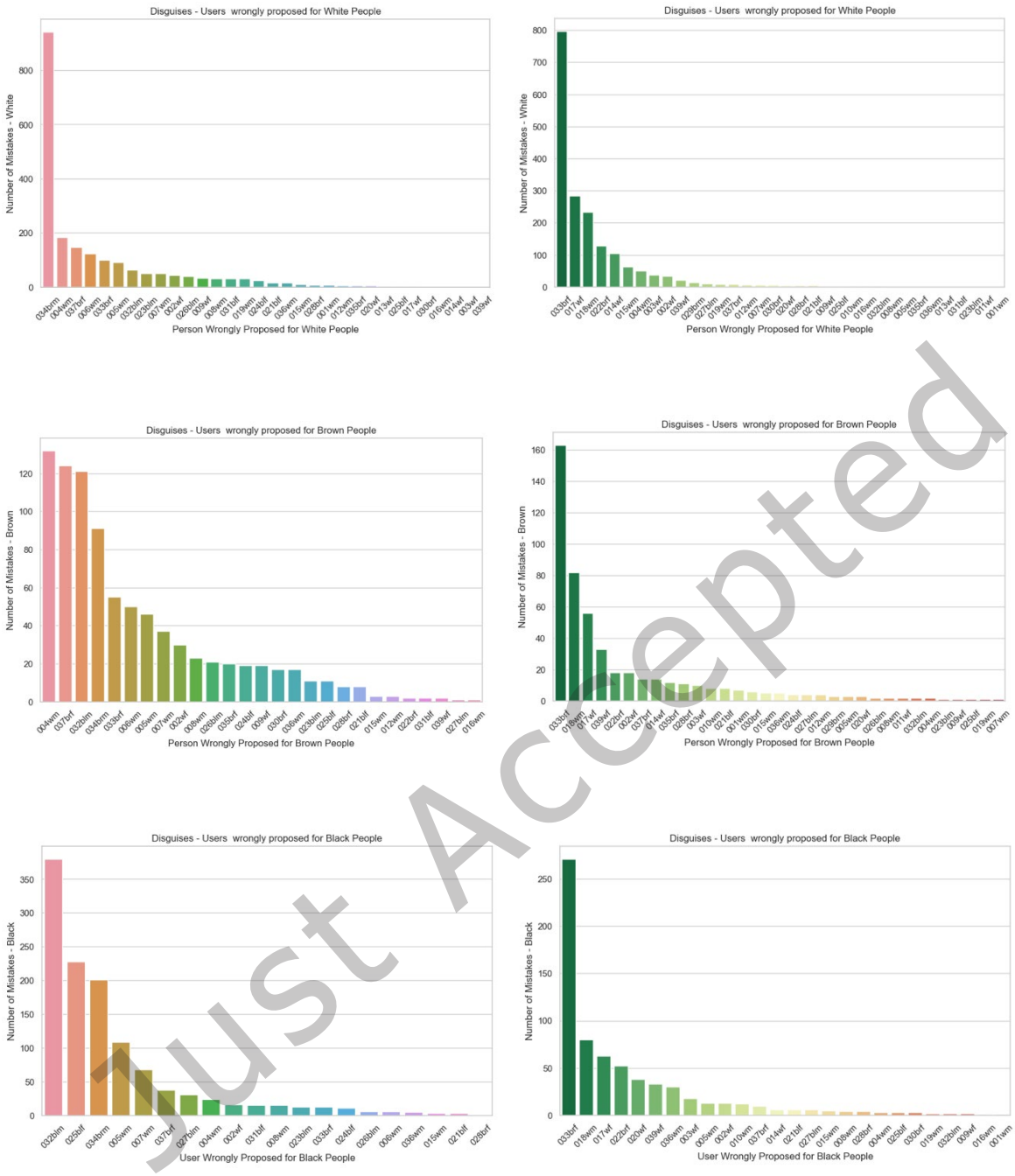


Fig. 19 A comparison of the misclassification for the three ethnicities from 2023 (left) and 2024 (right).

Table 14. The coefficients (coef), p-value (P>|z|) and [0.025, 0.977] show a 95% confidence interval (CI) of the logistic regression model, with the coefficients for disguise being compared to *clean* as reference level (one-hot encoding), the one for *gender* is compared to *female* and *ethnicity* compared to *black*.

Feature	Coefficient	P>|z|	[0.025	0.975]
Intercept	1.43	0.00	1.23	1.63
Gender[Male]	-0.04	0.439	-0.13	0.06

Ethnicity[Brown]	0.77	0.00	0.63	0.91
Ethnicity[White]	-0.53	0.00	-0.65	-0.42
Disguise[glasses_off]	-0.34	0.05	-0.58	-0.10
Disguise[glasses_on]	-0.74	0.00	-0.97	-0.50
Disguise[goggles]	-1.38	0.00	-1.61	-1.15
Disguise[halohat_off]	-0.34	0.05	-0.58	-0.10
Disguise[halohat_on]	-1.22	0.00	-1.45	-1.00
Disguise[rainbowhat_off]	-0.40	0.01	-0.64	-0.16
Disguise[rainbowhat_on]	-0.80	0.00	-1.03	-0.57
Disguise[snood_blobs]	-0.99	0.00	-1.23	-0.76
Disguise[snood_crosses]	-0.44	0.00	-0.68	-0.20
Disguise[snood_lines]	-1.44	0.00	-1.67	-1.21

Ethnicity also showed a significant impact. Compared to the black reference group, white ethnicity had a negative impact on performance, while brown ethnicity showed a positive impact. In contrast to the old model analyzed in section 5.2, *gender* did not demonstrate a significant impact on prediction capacity, with a p-value of 0.44. This suggests that *gender:male*, relative to the female reference group, did not significantly affect the FRT pipeline's predictions.

1.1.2 5.4.3 Verifying Model Consistency: A Comprehensive Re-evaluation

Up to now, our experiments occurred in a 'closed world' where the participants to be recognised where the same as those in the database of identities. In this section we report two experiments to assess our results against two different 'open worlds' where the database of identities holds a much larger set of people than the 39 individuals in our sample to be recognised in the recorded videos, this a more realistic settings in surveillance. To check the validity of our results in an 'open world' we conducted a comprehensive re-evaluation using an expanded face database against an increased frame sampling of the recorded videos. We first expanded the face matching database (database of identities) from the 39 participants to 2 databases containing 1000 and 3000 individuals randomly chosen within the LFW dataset[13]. We then compared the performance of the models with both the original and expanded databases. The combinations of MTCNN, QMagFace, and Euclidean models, which exhibited strong performance in previous benchmarks, were selected for this analysis. As there was no statistical significance difference between the 1000 and 3000 databases, we report only the result with the latter. Tables 15, 16, 17 show 3 main key findings:

- Increasing the identity database size from 39 to 3000 did not significantly affect the recognition of clean faces across all ethnicities.
- Certain disguises, particularly *goggles*, *snood_lines*, *snood_blobs*, and *snood_crosses*, had a more pronounced effect on recognition accuracy for *white* individuals. *Brown* individuals were also impacted by *goggles*.
- In some cases, expanding the database led to improved accuracy due to majority voting. For example, if a person's identity is initially ambiguous, adding more faces to the database could clarify the match and increase recognition accuracy. Consider a video of Luigi where 10 faces are extracted. With a database of 39 people, the system recognizes Luigi 4 times, Mario 5 times, and 1 time an unknown person. Therefore, the recognized person will be Mario, which is incorrect. When the database increases to 3000 people, the

[13] LFW Labeled Faces in the Wild https://vis-www.cs.umass.edu/lfw/

system recognizes Luigi 4 times, Mario 3 times, and 3 times a new person added to the database. Thus, through majority voting, the system correctly recognizes Luigi.

After this analysis, we also revisited previous experiments by increasing the number of sampled frames from the videos. This allowed us to determine if a larger number of frames would improve or degrade model performance. By sampling at 5 and 10 frames per second, we determined that the performance of our models was not statistically significantly affected. This confirms the reliability of our findings from the initial 2 frames per second sampling rate.

Table 15. Effectiveness of the disguises for white participants using one of the best recent models evaluated in the previous chapter expanding face database adding 3000 new faces provided by LFW database.

39 persons in the database *white: disguises*

Rank	System configuration	T	C	A	Disguise
clean	QMagFace : euclidean : Mtcnn	394	28	1.0	clean
1	QMagFace : euclidean : Mtcnn	345	21	0.95	snood_crosses
2	QMagFace : euclidean : Mtcnn	388	24	0.86	halohat_off
3	QMagFace : euclidean : Mtcnn	372	22	0.82	glasses_off
4	QMagFace : euclidean : Mtcnn	339	20	0.76	rainbowhat_off
5	QMagFace : euclidean : Mtcnn	351	18	0.71	snood_blobs
6	QMagFace : euclidean : Mtcnn	361	15	0.62	snood_lines
7	QMagFace : euclidean : Mtcnn	424	20	0.57	raimbowhat_on
8	QMagFace : euclidean : Mtcnn	357	18	0.55	glasses_on
9	QMagFace : euclidean : Mtcnn	365	15	0.5	goggles
10	QMagFace : euclidean : Mtcnn	428	154	0.32	halohat_on

3000 persons in the database *white: disguises*

Rank	System configuration	T	C	A
clean	QMagFace : euclidean : Mtcnn	373	247	0.95
4	QMagFace : euclidean : Mtcnn	325	147	0.68
1	QMagFace : euclidean : Mtcnn	366	208	0.9
3	QMagFace : euclidean : Mtcnn	356	183	0.81
2	QMagFace : euclidean : Mtcnn	323	165	0.85
7	QMagFace : euclidean : Mtcnn	338	105	0.3
10	QMagFace : euclidean : Mtcnn	345	74	0.2
5	QMagFace : euclidean : Mtcnn	406	171	0.65
6	QMagFace : euclidean : Mtcnn	336	153	0.57
8	QMagFace : euclidean : Mtcnn	343	84	0.24
9	QMagFace : euclidean : Mtcnn	408	109	0.24

Table 16. Effectiveness of the disguises for brown participants using one of the best recent models evaluated in the previous chapter expanding face database adding 3000 new faces provided by LFW database

39 persons in the database *brown: disguises*

Rank	System configuration	T	C	A	Disguise
clean	QMagFace : euclidean : Mtcnn	100	17	1.0	clean
1	QMagFace : euclidean : Mtcnn	211	17	1.0	glasses_on
2	QMagFace : euclidean : Mtcnn	186	15	1.0	halohat_off
3	QMagFace : euclidean : Mtcnn	196	17	1.0	glasses_off
4	QMagFace : euclidean : Mtcnn	181	15	1.0	rainbowhat_off
5	QMagFace : euclidean : Mtcnn	235	18	1.0	raimbowhat_on
6	QMagFace : euclidean : Mtcnn	201	14	1.0	snood_crosses
7	QMagFace : euclidean : Mtcnn	201	14	0.89	halohat_on
8	QMagFace : euclidean : Mtcnn	215	15	0.78	snood_lines
9	QMagFace : euclidean : Mtcnn	197	15	0.78	snood_blobs
10	QMagFace : euclidean : Mtcnn	228	149	0.67	goggles

3000 persons in the database *brown: disguises*

Rank	System configuration	T	C	A
clean	QMagFace : euclidean : Mtcnn	200	170	1.0
1	QMagFace : euclidean : Mtcnn	211	162	1.0
5	QMagFace : euclidean : Mtcnn	186	144	1.0
2	QMagFace : euclidean : Mtcnn	196	161	1.0
3	QMagFace : euclidean : Mtcnn	156	116	1.0
4	QMagFace : euclidean : Mtcnn	191	144	1.0
8	QMagFace : euclidean : Mtcnn	173	95	0.75
6	QMagFace : euclidean : Mtcnn	167	104	0.89
7	QMagFace : euclidean : Mtcnn	163	99	0.75
9	QMagFace : euclidean : Mtcnn	176	96	0.75
10	QMagFace : euclidean : Mtcnn	228	34	0.11

Table 17. Effectiveness of the disguises for black participants using one of the best recent models evaluated in the previous chapter expanding face database adding 3000 new faces provided by LFW database

39 persons in the database *black*: disguises

Rank	System configuration	T	C	A	Disguise

3000 persons in the database *black*: disguise

Rank	System configuration	T	C	A

clean	QMagFace : euclidean : Mtcnn	159	12	0.88	clean	clean	QMagFace : euclidean : Mtcnn	159	120	0.88
1	QMagFace : euclidean : Mtcnn	197	15	0.88	glasses_off	1	QMagFace : euclidean : Mtcnn	197	149	0.88
2	QMagFace : euclidean : Mtcnn	186	14	0.88	snood_crosses	7	QMagFace : euclidean : Mtcnn	186	118	0.75
3	QMagFace : euclidean : Mtcnn	208	15	0.75	glasses_on	3	QMagFace : euclidean : Mtcnn	208	144	0.75
4	QMagFace : euclidean : Mtcnn	176	13	0.75	halohat_off	2	QMagFace : euclidean : Mtcnn	176	135	0.88
5	QMagFace : euclidean : Mtcnn	200	14	0.75	halohat_on	4	QMagFace : euclidean : Mtcnn	200	140	0.75
6	QMagFace : euclidean : Mtcnn	172	13	0.75	raimbowhat_off	5	QMagFace : euclidean : Mtcnn	172	130	0.75
7	QMagFace : euclidean : Mtcnn	191	14	0.75	raimbowhat_o	6	QMagFace : euclidean : Mtcnn	191	138	0.75
8	QMagFace : euclidean : Mtcnn	196	92	0.38	goggles	9	QMagFace : euclidean : Mtcnn	196	34	0.12
9	QMagFace : euclidean : Mtcnn	175	76	0.38	snood_blobs	8	QMagFace : euclidean : Mtcnn	175	57	0.38
10	QMagFace : euclidean : Mtcnn	196	44	0.0	snood_lines	10	QMagFace : euclidean : Mtcnn	196	39	0.12

5.5 Lessons learnt from the evaluations

The three evaluations with participants and the additional cross tests inform our research in different ways. The first (5.3.1) showed slightly different accuracies by the two FRT (Facenet512 and FGGface) used with their performance in the *corridor* and *hall* settings poorer than in *environment: gate*. The change of the illumination along the corridor and the skewed top-down angle of view in the hall may explain the reduced performance and a more powerful camera may produce improved results. The second more extended evaluation (5.3.2) tested the easiest *environment: gate* well-lit front-face video capturing with a larger group of participants from different ethnicities. Even in this best-condition setting there is a high variation of results for the same system with different configurations. Moreover, different configurations perform better with different ethnicities making it impossible to select the best one overall. The last evaluation (5.3.3) assessed the advancements of FRT showing there has been substantial improvement but, when disguises are used, misclassifications still occur. We have also seen a potential tendency towards a lower performance when the sample increases.

The disguises were all successful in reducing the accuracy across all ethnicities albeit at different degrees. The two hats projecting light on the face of black participants made their faces easier to recognise thus having the opposite effect than intended. Of concern is the number of misrecognitions that occur in the clean and increase substantially with the disguises. It is worth noticing that misrecognitions occur across ethnicities as well as within ethnicities. It is then essential that the process of identification is supervised by a human who takes responsibility for the final decision[14].

Finally, we consider the design decision taken at the time of moving from the proof of concepts to the prototype. To guarantee wearability and ease of use during the evaluation we had to compromise some aspects of the design of the glasses (the light emitted, Fig. 6) and the rainbow hat (the pattern projected, Fig. 10). While the evaluation shows a drop in accuracy, this has been much less than we expected following the early tests. A design iteration to make prototypes that are closer to the initial concepts may produce the high deception expected.

6 LIMITATIONS

Our study has some limitations. First is the number of participants (39) that resulted in a very small number of faces tested when compared with large datasets of still images. However, a clear distinction and novelty of our work is the use of live video rather than still images: with 39 people of different ethnicities, our study is the largest of presentation attacks in real-life video-feed conditions.

[14] The identification procedure may vary from country to country. In the UK a police-person checks the FRT output and may decide to take the process of identification forward or may decide not to proceed. /https://www.met.police.uk/SysSiteAssets/media/downloads/central/services/accessing-information/facial-recognition/met-evaluation-report.pdf

A second limitation is the different size of groups belonging to different ethnicities with white participants being more than black and brown combined (56% vs. 44%). However, when compared with other studies in the literature, our sample has a more participants [Sharif et al. 2019] and a more balanced proportion [Rosenberg et al. 2023], a possible reason for achieving different findings. This shows the complexity in studying FRT.

A further limitation is due to the choices made when designing the final disguises. For reasons to do mostly with the comfort of participants, the final disguises implemented favoured ease of wear and use rather than technical solutions. Our choices, e.g. to project normal light rather than IR or UV, impacted on the ability of the disguises to bypass FRT, a negative impact much stronger than what we expected. We can only speculate that, had we used IR (infrared) or UV (ultra-violet) light the disguises would have been much more effective given the excellent FRT bypass achieved in preliminary tests (Fig. 3, 6, 9). In hindsight, focusing on the product rather than the technology was a mistake.

Following on the same attitude of focussing on the participants, we decided not to use commercial software as to secure the privacy of our volunteers (an example of 'privacy by design' [Almeida et al. 2022]). Therefore, our comparison across models and over time is limited to open access software we could control in full.

In our study we tried to replicate as close as possible realistic conditions. As such we should have used a CCTV system where the camera has been designed for indoor monitoring. Our attempt to use such devices was frustrated by CCTVs being close systems meaning that only compressed video is fed out as record. We considered compressed video not suitable as the performance of the FRT would have been lower not for the effectiveness of the disguises, but because of the video compression. Reports of the actual use of FRT in public spaces [Fussey and Murray 2019; Mansfield 2023] do not state at which point in the pipeline FRT is placed, from video acquisition to detection flagging, we can infer FRT is fed with the best quality images thus before the compression stage. Seen in the perspective of achieving a realistic simulation, our choice replicates the most likely pipeline but falls short of using the correct camera devices: this is a clear limitation albeit in line with other academic work.

Despite these limitations, our study contributes to shed some light on FRT when it is taken out of the lab and challenged, not through the expert knowledge of computer scientists that know how to trick it, but by the ingenuity of design.

7 CONCLUSIONS AND FUTURE WORK

Our work frames FRT through design practice and understands its use and impact through the lens of HCI. Our approach does not start from what the technology can or cannot do, rather from the question "if I did not want to be recognised, what would I do?" In a generative design process, first we imagined 120 disguises of different complexity and feasibility, then distilled those in 50 concepts to be tested in the lab as mock-ups to identify which ones successfully bypassed FRT. Our purpose was to empirically select the promising concepts to be prototyped for a wider evaluation. Since these early stages, a few of our concepts had similarities with published research, yet our lab tests (Fig. 3) were not always consistent with the literature: the synthetic makeup applied on photos in [Lin et al. 2022] deceived FRT while two of our makeups failed; face-on-face projection was successful in [Shen et al. 2019] while it was not in our experiment; a mask with a fake nose and mouth was successful in [Zolfi et al 2022] but failed out lab test. As [Lin et al. 2022], [Shen et al. 2019] and [Zolfi et al 2022] are all computational attacks, we may infer they were effective not because of the type of disguise (makeup, projection, mask), but because they have been maliciously crafted to target known FRT weaknesses. However, in our lab tests (Fig. 3) the geometric make-up and the pink-dots face projection successfully bypassed FRT showing that computational knowledge is not a 'must have'.

Our empirical exploration then progressed towards wearable prototypes that could be worn by participants in a comparative evaluation against different FRT settings. We selected the two ends of the spectrum: simple camouflage (the snoods) and complex interactives that projected light on the face (hats) or against the camera (glasses). At this stage of the design process our intent was to get closer to disguises that people (such as civil right activists) could wear in the street without raising alarm, disguises that could be easily put on and off (the snoods) or that could be switches on and off at will (the hats and the glasses). Moving from mock-ups to prototypes posed a challenge for the complex concepts: the UV and IR emitters attached to the glasses to attack the camera by projecting out light (Fig. 7) could not be seamlessly integrated in an off-the-shelf glass frame (the

lab mock-up was a larger 3D printed frame), and the pico-projector that shed a geometric pattern on the face (Fig. 10) could not be micronized to be integrated in a hat within the resources of the project. We then took the decision to favour wearability and to implement disguises that looked realistic to be used in the evaluation: the glasses had fibre optics, the hat an animated coloured led strip. Most likely we paid this choice with an increase in FRT detection (i.e., a decrease in the power of the disguises) as the disguises were not as effective as expected given the results of the lab test. In future work, it could be worth going back to the drawing board to find ways to incorporate the original concept into a wearable disguise, for example LEDs could be placed under the hat brim to create a sharp light-dark effect as in [Li et al. 2023] or infrared light projected used by [Zhou et al. 2018] although safety and comfort of the participants may be an issue here.

We evaluated the prototypes in three different settings that simulated real-world conditions: the cameras were placed in different spaces (a hall, a corridor, a passport gate); at different heights, angles and lighting conditions; the attacker was wearing a physical disguise and free to move as they liked. In the hall and the corridor FRT performed poorly respect to the front-face at the gate (Fig. 14). We could not find any study that attempted to replicate real environmental conditions or the behaviour of the attacker as we did. The standard FRT experimental material is one or more photos taken front-face of a person sat in front of a camera to which digital patches are automatically applied [Zhu et al. 2019, Pautov et al. 2019, Zolfi et al. 2022], sometimes multiple photos with the head in different positions and light variation [Komkov and Petiushko 2021, Lin et al. 2021], and in-person attacks wearing a disguise are rare (makeup in [Chen et al. 2019], glasses in [Sharif et al. 2016]). Our results in the hall and corridor suggests that the experimental conditions in which FRT has been evaluated so far do not reflect realistic human-centred scenarios. Taking FRT out of the lab poses new challenges to the research community and open up new avenues of investigation.

In the second evaluation we expanded the range of participants to include black and brown ethnicities. The findings show that disguises should be designed with a specific ethnicity in mind: *halohat_on* is very effective on *white* people as it 'cancels' some facial features (the nose and the face contour) while it makes *black* people much easier to recognise as it illuminates their faces.

Our analysis shows some individuals are easier to identify than others and that cases of misidentification can occur both within and across ethnicities. This last finding is the opposite of [Rosenberg et al. 2023] that only finds misidentifications occurring within ethnic group when testing FRT against digital obfuscation. However, in [Rosenberg et al. 2023] the number of people identified as of Indian ethnicity were only 20 against 1240 Black, 2.500 Asian and 10.000 White, while our sample was much smaller but more balanced between 22 White, 8 Black and 9 Brown (Indian ethnicity). This observation suggests the need to experiment with more balanced samples representing different ethnicities equally and that generalisations cannot be done as unbalanced data may lead to unreliable results. Indeed, in our experiments the most misrecognised individuals as *white* and as *black* in both 2023 and 2024 were *brown* people (Fig. 19), an ethnicity marginal in [Rosenberg et al. 2023] respect to the other groups.

Another explanation for the opposed results between our work and [Rosenberg et al. 2023] could be the different setup: as [Rosenberg et al. 2023] applied digital obfuscation and did not test physical presentation attacks, we may presume the disguises facilitated cross-ethnicities misidentifications in different ways. As explained in 3.1, FRT extracts the facial features and use them to position a point in a vector space expecting that to be the closest to the point of the actual face (ground truth) of the individual detected. When 'noise' is introduced, being that via digital obfuscation or physical disguises, some facial features are hidden or distorted forcing the FRT to interpret the facial features as a point positioned in the vector space far from the actual individual and closer to others resulting in the misidentification or the overlooking of the target person. It could be that the 'noise' from digital obfuscation and physical disguises occur on different features thus confound FRT in different ways. As the two studies differ in the evaluation environment, a real-life in our case vs. the LFW dataset (Labelled Faces in the Wild) for [Rosenberg et al. 2023], an interesting experiment could be to apply both in sequence to see how far FRT could be bypassed when noise is maximised.

It is also impossible to compare our results against other physical presentation attacks as only [Sharif et al. 2019] included more ethnicities, 1 south Asian female and 1 middle eastern man as impersonators. The lack of comparable work shows there are many opportunities to research FRT from a human-centred real-life

perspective and that FRT research would benefit from empirical findings, the involvement of more than a few participants of different ethnicities, all factors that could bring lab research closer to real uses.

The cases of misidentification are particularly critical for the potential consequences this could bring: [Mansfield 2023] reports that in a real-life evaluation of commercial FRT used by the UK police the cases of misidentification (false positive) where 1 in 60,000 with a watchlist on 1000 individuals but increased to 1 in 6000 when the watchlist increases to 10,000 (with 89% true positive for both settings). The consequences of FRT misidentification if not properly verified and confirmed by a human are serious [BBC 2019, Hill 2020, Hill 2022, Murphy 2023] and the number of people potentially affected when scaling up to the population of an entire country raise procedural and ethical questions even more when the technology is used by the private sector, for example when attempting to identify shoplifters [Chivers 2019]. The need to regulate FRT and its use would be a start, better would be to impose "privacy by design" and "privacy by default" to cover the creation, processing, sharing, and destruction of personal data as in the EU's General Data Protection Regulation (GDPR) [Almeida et al. 2022] and to avoid abuses [REUTERS 2023]. While GDPR does not prevent data collection, it requires carefully documented processes and data management, but "what is considered fair and lawful is potentially open to interpretation" [Almeida et al. 2022, pg. 380] leaving open to individual countries how the government controls FRT and who is ultimately responsible for its use.

The research in this paper was carried out over a period of 3 years, from the design and lab testing of the disguises in Spring 2021, to prototyping and real-life evaluation in three different settings, to the final comparative evaluation in Winter 2023-24. During this period, we have experimented with different models keeping the study up-to-date with the latest FR developments. While there is an improvement with recent models, when disguises are in use, misidentification still occurs: the relationship between the disguise, specific facial features and misidentification is worth exploring further.

ACKNOWLEDGMENTS

The authors thank Ursula Ankeny and Josh Hill for their contribution to the design of the disguises and the assistance during the experiments.

REFERENCES

Alessandro Acquisti, Ralph Gross and Fred Stutzman. 2014. Face Recognition and Privacy in the Age of Augmented Reality. Journal of Privacy and Confidentiality. 6 (82). 1-20.

Denise Almeida, Konstantin Shmarko and Elizabeth Lomas. 2021. The Ethics of Facia Recognition Technologies, Surveillance, and Accountabilityin an Age of Artificial Intelligence: A Comparative Analysis of US, UE, and UK Regulatory Frameworks. AI and Ethics. 2. 377-387.

Paarijaat Aditya, Rijurekha Sen, Peter Druschel, Seong Joon Oh, Rodrigo Benenson, Mario Fritz, Bernt Schiele, Bobby Bhattacharjee and Tong Tong Wu. 2016. I-Pic: A Platform for Privacy-Compliant Image Capture. Proc. of ACM MobiSys'16. June 25-30, Singapore. DOI: 10.1145/2906388.2906412

Mohamad Alansari, Oussama Abdul Hay, Sajid Javed, Abdulhadi Shoufan, Yahya Zweiri and Werghi, Naoufel. 2023. GhostFaceNets: Lightweight Face Recognition Model from Cheap Operations. IEEE Access. PP. 1-1. 10.1109/ACCESS.2023.32660688

BBC. 2019. Apple AI accused of leading to man's wrongful arrest. BBC News. 23 April 2019. Retrieved from https://www.bbc.co.uk/news/technology-48022890

Johana Bhuiyan, 2023. Rite Aid facial recognition misidentified Black, Latino and Asia people as 'likely' shoplifter. The Guardian. 20 December 2023. Retrieved from https://www.theguardian.com/technology/2023/dec/20/rite-aid-shoplifting-facial-recognition-ftc-settlement

Johana Bhuiyan, 2024. Facial recognition used after Sunglass Hut robbery led to man's wrongfully jailing, say suit. The Guardian. 23 January 2024. Retrieved from https://www.theguardian.com/technology/2024/jan/22/sunglass-hut-facial-recognition-wrongful-arrest-lawsuit

Carmen Bisogni, Lucia Cascone, Jean-Luc Dugelay and Chiara Pero. 2021. Adversarial Attacks through Architectures and Spectra in Face Recognition. Pattern Recognition Letters. 147. 55-62. https://doi.org/10.1016/j.patrec.2021.04.004

Fadi Boutros, Naser Damer, Florian Kirchbuchner, Arjan Kuijper; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022, pp. 1578-1587

Fadi Boutros, Vitomir Struc, Julian Fierrez and Naser Damer. 2023. Synthetic data for face recognition: Current state and future prospects. Image and Vision Computing. Vol. 135. DOI: https://doi.org/10.1016/j.imavis.2023.104688

Tim Brown. 2009. Change by Design: How Design Thinking Transforms Organisations and Inspires Innovation. Harper Collins.

Qingxiu Bu. 2021. The global governance on automated facial recognition (AFR): ethical and legal opportunities and privacy challenges. International Cybersecurity Law Review. 2:113-145.

Joy Buolamwini and Tinit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proc. of Machine Learning Research. Conference on Fairness, Accountability, and Transparency. 81:1-15.

Davide Castelvecchi. 2020. Beating Biometric Bias. Nature, 587, 19 November 2020, 347-349.

Cunjian Chen, Cunjian Dantcheva, Thomas Swearingen and Ann Ross. 2017. Spoofing Faces Using Makeup: An Investigative Study. Proc. of 3rd IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2017), (New Delhi, India), February 2017.

Tom Chivers. 2019. Facial recognition… coming to a supermarket near you. The Guardian, 4 August 2019. Retrieved from https://www.theguardian.com/technology/2019/aug/04/facial-recognition-supermarket-facewatch-ai-artificial-intelligence-civil-liberties

Debayan Deb, Xiaoming Liu and Anil K. Jain. 2023. Unified Detection of Digital and Physical Face Attacks. Proc. of IEEE 17th International Conference on Automatic Face and Gesture Recognition. DOI: 10.1109/FG57933.2023.10042500

Peter Fussey and Daragh Murray. 2019. Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology. Project Report. University of Essex Human Rights Centre. Retrieved from https://repository.essex.ac.uk/24946/

Rakibul Hasan, David Crandall, Mario Fritz and Apu Kapadia. 2020. Automatically Detecting Bystanders in Photos to Reduce Privacy Risks. Proc. of IEEE Symposium on Security and Privacy. 18-21 May, San Francisco, CA, USA. DOI: 10.1109/SP40000.2020.00097

Amy Hawkins. 2018. Beijing's Big Brother Tech Needs African Faces. Foreign Policy. July 24 2018. Retrieved from https://foreignpolicy.com/2018/07/24/beijings-big-brother-tech-needs-african-faces/

Kashmir Hill. 2023. The Secretive Company That Might End Privacy as We Know It. The New York Times. International Edition. January 6, 2023.

Kashmir Hill. 2020. Eight Months Pregnant Woman Arrested After False Facial Recognition Match. The New York Times. August 20, 2020.

Javier Hernandez-Ortega, Julian Fierrez, Aythami Morales and Javier Galbally. 2019. Introduction to Face Presentation Attack Detection. In: Marcel, S., Nixon, M., Fierrez, J., Evans, N. (eds.) Handbook of Biometric Anti-Spoofing – Advances in Computer Vision and Pattern Recognition. Springer Cham. 187-206. DOI: https://doi.org/10.1007/978-3-319-92627-8_9

Siddarth Jaiswal, Karthikeya Duggirala, Abhisek Dash and Animesh Mukherjee. 2022. Two-Face: Adversarial Audit of commercial Face Recognition Systems. Proc. of 16th International AAAI Conference on Web and Social Media (ICWSM 2022). 381-392.

Stepan Komkov and Aleksandr Petiushko. 2021. AdvHat: Real-World Adversarial Attack on ArcFace ID System. Proc. of 25th International Conference on Pattern Recognition (IPCR 2021). DOI: 10.1109/ICPR48806.2021.9412236

Yassin Kortli, Maher Jtdi, Ayman Al Falou and Mohamed Arti. 2020. Face Recognition System: A Survey. Sensors 20, 342. DOI: https://doi.org/10.3390/s20020342

Ilpo Koskinen, John Zimmerman, Tomas Binder, Johan Redström and Stephan Wensveen. 2011. Design Research Through Practice: From the Lab, Field and Showroom. Morgan Kaufmann.

Yanjie Li, Yiquan Li, Xuelong Dai, Songtao Guo and Bin Xiao. 2023. Physical-World Optical Adversarial Attacks on 3D Face Recognition. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 24699-24708, doi: 10.1109/CVPR52729.2023.02366.

Chang-Sheng Lin, Chia-Yi Hsu, Pin-Yu Chen and Chia-Mu Yu. 2022. Real-World Adversarial Examples via Makeup. Proc. of IEEE International conference on Acustic, Speech and Signal Processing (ICASSP 2022). 2854-2858 DOI: 10.1109/ICASSP43922.97047469

Sascha Löbner, Sebastian Pape, Vanessa Bracamonte. 2023. User Acceptance Criteria for Privacy Preserving Machine Learning Techniques. 18th International Conference on Availability, Reliability and Security (ARES 2023). https://dl.acm.org/doi/fullHtml/10.1145/3600160.3605004

Tony Mansfield. 2023. Facial Recognition Technology in Law Enforcement Equitability Study (Final Report). NPL Report MS 43. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://science.police.uk/site/assets/files/3396/frt-equitability-study_mar2023.pdf

Tom Murphy. 2023. Rite Aid Banned from Facial Recognition Tech Use for 5 years After Faulty Theft Targeting in Store. Time. 20 December 2023. https://time.com/6549652/rite-aid-banned-facial-recognition-tech/

Dinh-Luan Nguyen, Sunpreet S. Arora, Yuhang Wu, Hao Yang. 2020. Adversarial Light Projection Attack on Face Recognition Systems: A Feasibility Study. IEEE Computer Vision and Pattern Recognition (CVPR) Biometrics Workshop 2020. Accessible at https://arxiv.org/abs/2003.11145

Seong Joon Oh, Rodrigo Benenson, Mario Fritz and Bert Schiele. 2016. Faceless Person Recognition; Privacy Implications in Social Media. Proc. of European Conference in Computer Vision ECCV 2026, pp. 19-35, Spinger Lecture Notes in Computer Science LNIP vol. 9907. DOI: 10.1007/978-3-319-46487-9_2

Mikhail Pautov, Grigorii Melnikov, Edgar Kaziakhmedov, Klim Kireev and Aleksandr Petiushko. 2019. On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System. 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) DOI: https://doi.org/10.1109%2Fsibircon48586.2019.8958134

Alina Polyakova and Chris Meserole. 2019. Exporting digital authoritarianism. Brookings Foreign Policy. Democracy and Disorder. https://www.brookings.edu/wp-content/uploads/2019/08/FP_20190826_digital_authoritarianism_polyakova_meserole.pdf

Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee and Emily Denton. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. AAAI/ACM Conference on AI, Ethics and Society - AIES '20, February 7–8, 2020, New York, NY, USA

Raghavendra Ramachandra and Christoph Busch. 2017. Presentation Attack Detection Methods for Face Recognition Systems: A

Comprehensive Survey. ACM Computing Surveys, 50, 1, Article 8. DOI: http://dx.doi.org/10.1145/3038924

REUTERS (2023) Privacy group challenges Ryanair's use of facial recognition. 27 July 2023 Retrieved from https://www.reuters.com/business/aerospace-defense/privacy-group-challenges-ryanairs-use-facial-recognition-2023-07-27/.

Harrison Rosenberg, Brian Tang and Somesh Jha. 2023. Fairness Properties of Face Recognition and Obfuscation Systems. Proc of 32nd USENIX Security Symposium. 7231-7248. https://doi.org/10.48550/arXiv.2108.02707

Aimee Kendall Roundtree. 2021. Ethics and Facial Recognition Technology: An Integrative Review. Proc. of 3rd World Symposium on Artificial Intelligence (WSAI 2021). IEEE Publisher. DOI: 10.1109/WSAI51899.2021.9486382

CO, USA. DOI: https://doi.org/10.1145/3314111.3319913

Evan Selinger and Brenda Leong. 2021. The Ethics of Facial Recognition Technology. In: C. Véliz (ed.) The Oxford Handbook of Digital Ethics. Oxford University Press. DOI: 10.2139/ssrn.3762185 Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3762185

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer and Michael k. Reiter. 2016. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. ACM Computer and Communication Security CCS'16. 1528-1540.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer and Michael k. Reiter. 2019. A General Framework for Adversarial Examples with Objectives. ACM Trans. Priv. Secur. 22, 3, Article 16. DOI: https://doi.org/10.1145/3317611

Emine Sinmaz. 2023. Live facial recognition labelled 'Orwellinan' as Met police push ahead with use. The Guardian. 5 April 2023. Retrieved from https://www.theguardian.com/technology/2023/apr/05/live-facial-recognition-criticised-metropolitan-police

Meng Shen, Zelin Liao, Liewhuang Zhu, Ke Xu and Xiaojiang Du. 2019. VLA: A Practical Visible Light-based Attack on Face Recognition Systems in Physical World. Proc of ACM Interactive, Mobile, Wearable and Ubiquitous Technologies, 3, 3, Article No: 103. https://doi.org/10.1145/3351261

Pieter Jan Stappers. 2007. Doing Design as a Part of doing Research. In: R. Michel. Design Research Now – Essays and Selected Projects.

Julian Steil, Marion Koelle, Wilko Heuten, Susanne Bolol and Andreas Bulling. 2019. PrivacEye: Privacy-Preserving Head-Mounted Eye Tracking Using Egocentric Scene Image and Eye Movement Features. Proc. of Eye Tracking Research and Applications (ETRA'19), June 25-28, Denver.

Philipp Terhörst, Malte Ihlefeld, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran Raja and Arjan Kuijper. 2023. QMagFace: Simple and Accurate Quality-Aware Face Recognition. 3473-3483. 10.1109/WACV56688.2023.00348

Fatemeh Vakhshiteh, Ahmad Nickabadi and Raghavendra Ramachandra. 2021. Adversarial Attacks Against Face Recognition: A comprehensive Study. IEEE Access. DOI: https://doi.org/10.1109/ACCESS.2021.3092646

Richard Van Noorden. 2020. The Ethical Questions that Haunt Facial-Recognition Research. Nature. Vol. 587. (19 November 2020), 354-358.

Changsheng Wan, Li Wang, and Vir V. Phoha. 2018. A Survey on Gait Recognition. ACM Comput. Surv. 51, 5, Article 89 (September 2019), 35 pages. https://doi.org/10.1145/3230633

Robert Williams. 2020. I was Wrongly Arrested Because of Face Recognition. Why Are Police allowed to Use It? The Washington Post. June 24, 2020.

Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang and Xue Lin. 2020. Adversarial T-Shirt! Evading Person Detectors in a Physical World. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. Lecture Notes in Computer Science, 12350. Springer. https://doi.org/10.1007/978-3-030-58558-7_39

Jialiang Zhang, Lixiang Lin, Jianke Zhu, Yang Li, Yun-chen Chen and Yao Hu. 2021. Attribute-Aware Pedestrian Detection in a Crowd. in IEEE Transactions on Multimedia, vol. 23. 3085-3097. doi: 10.1109/TMM.2020.3020691

Shikun Zhang, Yuanyuan Feng, and Norman Sadeh. 2021. Facial Recognition: Understanding Privacy Concerns and Attitudes Across Increasingly Diverse Deployment Scenarios.16th Symposium on Usable Privacy and Security. August9-10. https://www.usenix.org/system/files/soups2021-zhang-shikun.pdfZhe Zhou, Di Tang, Xiaofeng Wang, Weili Han, Xiangyu Liu and Kehuan Zhang. 2018. Invisible Mask: Practical Attacks on Face Recognition with Infrared. Retrieved from ARXIV repository (Cornell University) https://doi.org/10.48550/arXiv.1803.04683

Zheng-Han Zhu, Yun-Zhong Lu, Chen-Kuo Chiang. 2019 Generating Adversarial Examples By Makeup Attacks on Face Recognition, IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 2516-2520, doi: 10.1109/ICIP.2019.8803269.

Alon Zolfi, Shai Avidan, Yuval Elovici and Asaf Shabtai. 2022. Adversarial Mask: Real-World Adversarial Attack Against Face Recognition Models. 22nd Joint European conference on Machine Learning and Principles and Practice of Knowledge discovery in Databases – ECML PKDD 2022. Springer Lecture Notes in Computer Science. DOI: https://doi.org/10.1007/978-3-031-26409-2_19

APPENDICES

A.1 Tables of the top results for *disguise:clean*

ethnicity:white

Model	Distance_Metric	Backend	Count	Total	Accuracy
Facenet512	euclidean_l2	retinaface	242	391	1.0
Facenet512	euclidean	retinaface	225	414	0.95
ArcFace	cosine	retinaface	185	367	0.77
ArcFace	euclidean_l2	retinaface	184	361	0.77
Facenet512	euclidean	opencv	202	465	0.77
Facenet512	euclidean_l2	opencv	221	448	0.77
ArcFace	euclidean	retinaface	172	374	0.77
Facenet	cosine	retinaface	164	329	0.73
Facenet512	cosine	retinaface	126	222	0.73
VGG-Face	cosine	retinaface	200	409	0.73
VGG-Face	euclidean_l2	retinaface	200	391	0.73
VGG-Face	cosine	opencv	176	430	0.68
VGG-Face	euclidean	retinaface	161	422	0.68
Facenet	euclidean_l2	retinaface	135	296	0.68
Facenet	euclidean	retinaface	142	329	0.68
VGG-Face	euclidean_l2	opencv	175	420	0.68
ArcFace	cosine	opencv	163	418	0.59
ArcFace	euclidean_l2	opencv	161	413	0.59
ArcFace	euclidean	opencv	143	409	0.55
Facenet	cosine	opencv	141	383	0.55
DeepFace	euclidean_l2	retinaface	133	412	0.5
DeepFace	euclidean	ssd	94	282	0.5
Facenet512	cosine	opencv	109	264	0.5
VGG-Face	euclidean	opencv	152	469	0.5
DeepFace	cosine	retinaface	164	546	0.45
Facenet	euclidean	opencv	115	369	0.41

ethnicity:brown

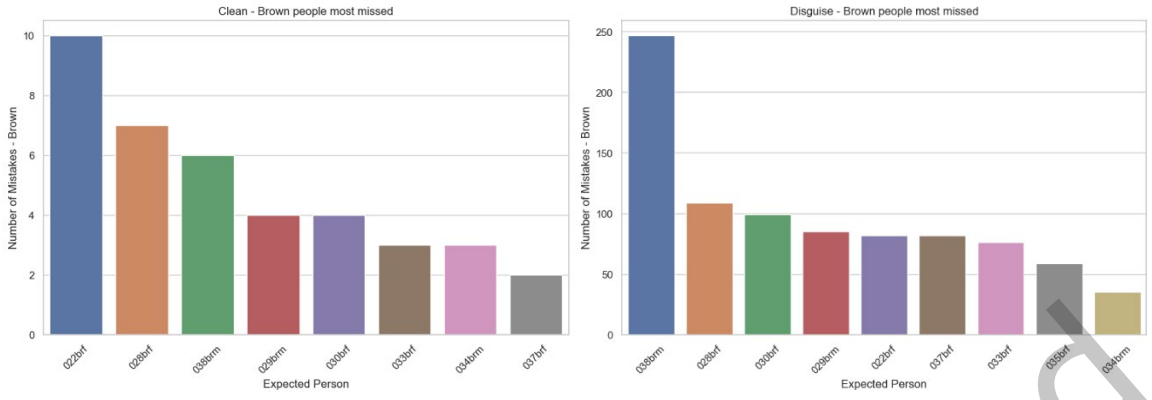
Model	Distance_Metric	Backend	Count	Total	Accuracy
Facenet	cosine	opencv	121	186	0.89
Facenet512	cosine	ssd	116	172	0.89
Facenet	euclidean	opencv	120	194	0.89
VGG-Face	euclidean_l2	ssd	121	203	0.89
VGG-Face	cosine	ssd	121	204	0.89
VGG-Face	euclidean_l2	retinaface	146	202	0.89
VGG-Face	euclidean	retinaface	130	206	0.89
VGG-Face	cosine	retinaface	147	204	0.89
VGG-Face	euclidean_l2	opencv	134	208	0.89
VGG-Face	euclidean	opencv	116	209	0.89
VGG-Face	cosine	opencv	134	208	0.89
Facenet512	euclidean	ssd	126	204	0.89
Facenet512	euclidean_l2	ssd	131	199	0.89
Facenet512	euclidean_l2	retinaface	158	197	0.89
Facenet512	euclidean	retinaface	154	204	0.89
Facenet512	cosine	retinaface	135	162	0.89
Facenet512	euclidean_l2	opencv	146	203	0.89
Facenet512	euclidean	opencv	142	210	0.89
Facenet512	cosine	opencv	122	165	0.89
Facenet	euclidean_l2	ssd	110	184	0.89
Facenet	euclidean	ssd	110	191	0.89
Facenet	cosine	ssd	112	191	0.89
Facenet	euclidean_l2	opencv	115	176	0.89
ArcFace	cosine	ssd	106	203	0.78
Facenet	euclidean_l2	retinaface	130	170	0.78
Facenet	euclidean	retinaface	131	184	0.78
Facenet	cosine	retinaface	133	178	0.78
ArcFace	euclidean_l2	ssd	105	201	0.78
ArcFace	euclidean	ssd	97	191	0.78

ethnicity:black

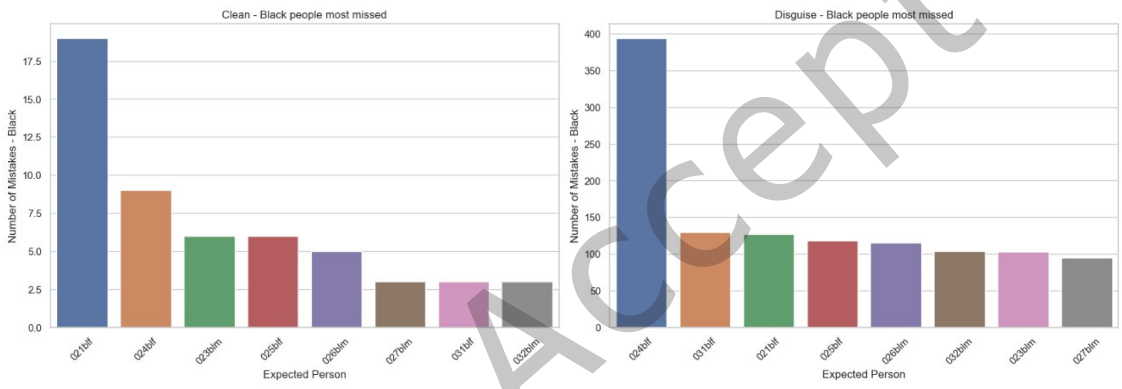
Model	Distance_Metric	Backend	Count	Total	Accuracy
Facenet512	euclidean	opencv	94	168	0.88
Facenet512	euclidean	retinaface	108	162	0.88
Facenet512	euclidean_l2	opencv	87	164	0.75
VGG-Face	euclidean_l2	retinaface	97	162	0.75
Facenet512	euclidean	ssd	87	160	0.75
Facenet512	euclidean_l2	retinaface	98	158	0.75
VGG-Face	cosine	retinaface	97	162	0.75
Facenet	cosine	opencv	72	152	0.75
ArcFace	cosine	ssd	77	156	0.75
ArcFace	euclidean_l2	retinaface	81	126	0.75
ArcFace	cosine	retinaface	87	142	0.75
Facenet512	cosine	retinaface	74	100	0.62
VGG-Face	euclidean_l2	ssd	79	160	0.62
VGG-Face	euclidean	ssd	65	160	0.62
VGG-Face	cosine	ssd	79	160	0.62
VGG-Face	euclidean	retinaface	63	162	0.62
VGG-Face	euclidean_l2	opencv	84	168	0.62
VGG-Face	cosine	opencv	84	168	0.62
Facenet512	euclidean_l2	ssd	82	159	0.62
Facenet	euclidean	opencv	73	152	0.62
Facenet	euclidean	retinaface	69	122	0.62
Facenet	cosine	retinaface	69	117	0.62
ArcFace	euclidean_l2	ssd	73	148	0.62
ArcFace	euclidean	ssd	69	147	0.62
ArcFace	euclidean	retinaface	72	126	0.62
ArcFace	euclidean_l2	opencv	85	158	0.62
ArcFace	cosine	opencv	87	161	0.62
Facenet	euclidean	ssd	65	146	0.5
VGG-Face	euclidean	opencv	71	168	0.5
Facenet512	cosine	ssd	71	124	0.5
Facenet512	cosine	opencv	51	104	0.5
Facenet	euclidean_l2	ssd	63	141	0.5
ArcFace	euclidean	opencv	72	147	0.5

A.2 Comparative charts clean vs. disguises for brown and black

Number of mistakes made in recognizing brown in the benchmark clean (left) and across all the disguises (right).



Number of mistakes made in recognizing black in the benchmark clean (left) and across all the disguises (right).



A.3 Performance of the 2 best configurations testing the extended database of identities

The primary database of identities uses a single high-quality photo taken full front on a white background to replicate a real situation where photos in passports, identity cards, driving licences, or mugshots are used to identify people. A second database was created adding to the 1 high-quality image a further 4 headshots taken from the clean video when the participant is in different positions. The tables below report the accuracy of the two best configurations, namely *Facenet512-euclidean-retinaface* and *Facenet512-euclidean_l2-retinaface*, on 39 participants with *environemnt:gate*.

disguise:clean ethnicity:white

Model	Distance_Metric	Backend	Count	Total	Accuracy
Facenet512	euclidean_l2	retinaface	194	369	0.73
Facenet512	euclidean	retinaface	169	399	0.64

disguise:clean ethnicity:brown

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
clean	Facenet512	euclidean	retinaface	92	203	0.67
clean	Facenet512	euclidean_l2	retinaface	104	203	0.67

disguise:clean ethnicity:black

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
clean	Facenet512	euclidean	retinaface	86	162	0.62
clean	Facenet512	euclidean_l2	retinaface	91	162	0.62

All disguises *ethnicity:white*

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
glasses_off	Facenet512	euclidean_l2	retinaface	123	351	0.64
halohat_off	Facenet512	euclidean_l2	retinaface	175	377	0.64
rainbowhat_off	Facenet512	euclidean_l2	retinaface	143	352	0.64
rainbowhat_on	Facenet512	euclidean_l2	retinaface	178	421	0.64
glasses_on	Facenet512	euclidean	retinaface	98	375	0.41
snood_crosses	Facenet512	euclidean_l2	retinaface	70	287	0.29
halohat_on	Facenet512	euclidean_l2	retinaface	101	356	0.27
goggles	Facenet512	euclidean_l2	retinaface	75	328	0.25
snood_blobs	Facenet512	euclidean_l2	retinaface	53	276	0.24
snood_lines	Facenet512	euclidean_l2	retinaface	33	298	0.14

All disguises *ethnicity:brown*

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
glasses_on	Facenet512	euclidean_l2	retinaface	136	233	0.89
glasses_off	Facenet512	euclidean	retinaface	110	227	0.67
rainbowhat_off	Facenet512	euclidean_l2	retinaface	100	197	0.67
goggles	Facenet512	euclidean_l2	retinaface	126	252	0.56
halohat_off	Facenet512	euclidean	retinaface	79	193	0.56
halohat_on	Facenet512	euclidean	retinaface	78	216	0.56
rainbowhat_on	Facenet512	euclidean	retinaface	89	266	0.33
snood_blobs	Facenet512	euclidean	retinaface	7	197	0.11
snood_crosses	Facenet512	euclidean	retinaface	25	218	0.11
snood_lines	Facenet512	euclidean	retinaface	30	239	0.11

All disguises *ethnicity:black*

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
rainbowhat_on	Facenet512	euclidean	retinaface	107	192	0.88
glasses_off	Facenet512	euclidean_l2	retinaface	124	199	0.75
glasses_on	Facenet512	euclidean	retinaface	127	211	0.75
halohat_off	Facenet512	euclidean	retinaface	111	179	0.75
rainbowhat_off	Facenet512	euclidean_l2	retinaface	100	174	0.75
halohat_on	Facenet512	euclidean	retinaface	98	203	0.62
goggles	Facenet512	euclidean_l2	retinaface	73	181	0.43
snood_crosses	Facenet512	euclidean	retinaface	27	186	0.25
snood_blobs	Facenet512	euclidean	retinaface	10	179	0.12
snood_lines	Facenet512	euclidean	retinaface	9	200	0.0

B.1 Evaluation 2024

The extended results of the many combinations of *model: metric : backend*

disguise:clean ethnicity:white

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
clean	QMagFace	euclidean	Mtcnn	282	394	1.0
clean	QMagFace	euclidean_l2	Mtcnn	284	394	1.0
clean	ElasticFace-Arc+	cosine	retinaface	245	403	0.95
clean	ElasticFace-Arc+	euclidean_l2	retinaface	245	403	0.95
clean	ElasticFace-Arc+	cosine	opencv	239	403	0.91
clean	ElasticFace-Arc+	euclidean_l2	opencv	239	403	0.91
clean	QMagFace	cosine	Mtcnn	270	394	0.91
clean	ElasticFace-Cos	cosine	retinaface	223	385	0.86
clean	ElasticFace-Cos	euclidean_l2	retinaface	223	385	0.86
clean	GhostFaceNet	euclidean_l2	retinaface	201	403	0.86
clean	ElasticFace-Arc+	euclidean	retinaface	216	403	0.82
clean	ElasticFace-Arc+	euclidean	opencv	217	403	0.77
clean	ElasticFace-Cos	cosine	opencv	211	385	0.76
clean	ElasticFace-Cos	euclidean_l2	opencv	211	385	0.76
clean	ElasticFace-Cos+	euclidean_l2	opencv	174	385	0.71
clean	ElasticFace-Cos+	cosine	opencv	174	385	0.71
clean	ElasticFace-Arc+	cosine	ssd	154	403	0.64
clean	ElasticFace-Arc+	euclidean_l2	ssd	154	403	0.64
clean	GhostFaceNet	euclidean_l2	opencv	178	403	0.64
clean	GhostFaceNet	cosine	retinaface	167	403	0.64
clean	GhostFaceNet	euclidean	retinaface	173	403	0.64
clean	ElasticFace-Cos	euclidean	opencv	197	385	0.62
clean	ElasticFace-Cos	euclidean	retinaface	191	385	0.57
clean	ElasticFace-Cos+	cosine	retinaface	172	385	0.57
clean	ElasticFace-Cos+	euclidean_l2	retinaface	172	385	0.57
clean	GhostFaceNet	cosine	opencv	156	403	0.45
clean	ElasticFace-Cos+	euclidean	opencv	151	385	0.43
clean	ElasticFace-Cos+	euclidean	retinaface	137	385	0.43
clean	GhostFaceNet	euclidean	opencv	147	403	0.41
clean	ElasticFace-Cos+	euclidean_l2	ssd	124	385	0.29
clean	ElasticFace-Cos+	cosine	ssd	124	385	0.29
clean	ElasticFace-Arc+	euclidean	ssd	148	403	0.18
clean	ElasticFace-Cos	cosine	ssd	143	385	0.14
clean	ElasticFace-Cos	euclidean_l2	ssd	143	385	0.14
clean	GhostFaceNet	cosine	ssd	88	403	0.14
clean	GhostFaceNet	euclidean_l2	ssd	111	403	0.14
clean	ElasticFace-Cos+	euclidean	ssd	99	385	0.1
clean	GhostFaceNet	euclidean	ssd	90	403	0.09
clean	ElasticFace-Cos	euclidean	ssd	138	385	0.05

disguise:clean ethnicity:brown

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
clean	ElasticFace-Arc+	cosine	retinaface	145	205	1.0
clean	QMagFace	cosine	Mtcnn	170	200	1.0
clean	QMagFace	euclidean	Mtcnn	174	200	1.0
clean	QMagFace	euclidean_l2	Mtcnn	177	200	1.0
clean	ElasticFace-Arc+	euclidean_l2	retinaface	145	205	1.0
clean	ElasticFace-Cos	euclidean_l2	opencv	134	205	0.89
clean	ElasticFace-Cos	euclidean_l2	retinaface	137	205	0.89
clean	ElasticFace-Cos	cosine	retinaface	137	205	0.89
clean	ElasticFace-Cos	cosine	opencv	134	205	0.89
clean	ElasticFace-Arc+	euclidean_l2	opencv	142	205	0.89
clean	ElasticFace-Arc+	euclidean	opencv	132	205	0.89
clean	ElasticFace-Arc+	cosine	opencv	142	205	0.89
clean	GhostFaceNet	cosine	retinaface	128	205	0.78
clean	GhostFaceNet	euclidean_l2	ssd	110	205	0.78
clean	GhostFaceNet	cosine	ssd	107	205	0.78
clean	GhostFaceNet	euclidean_l2	retinaface	134	205	0.78
clean	GhostFaceNet	euclidean	retinaface	132	205	0.78
clean	ElasticFace-Cos+	euclidean_l2	opencv	117	205	0.78
clean	ElasticFace-Cos+	cosine	opencv	117	205	0.78
clean	ElasticFace-Arc+	euclidean_l2	ssd	123	205	0.78
clean	ElasticFace-Arc+	cosine	ssd	123	205	0.78
clean	ElasticFace-Arc+	euclidean	retinaface	131	205	0.78
clean	ElasticFace-Arc+	euclidean	ssd	119	205	0.67
clean	ElasticFace-Cos+	cosine	retinaface	117	205	0.67
clean	ElasticFace-Cos+	euclidean	retinaface	106	205	0.67
clean	ElasticFace-Cos+	euclidean_l2	retinaface	117	205	0.67
clean	ElasticFace-Cos+	cosine	ssd	113	205	0.67
clean	ElasticFace-Cos+	euclidean	ssd	95	205	0.67
clean	ElasticFace-Cos+	euclidean_l2	ssd	113	205	0.67
clean	ElasticFace-Cos	euclidean_l2	ssd	120	205	0.56
clean	GhostFaceNet	euclidean	ssd	111	205	0.56
clean	ElasticFace-Cos+	euclidean	opencv	94	205	0.56
clean	ElasticFace-Cos	euclidean	retinaface	125	205	0.56
clean	ElasticFace-Cos	euclidean	ssd	119	205	0.56
clean	ElasticFace-Cos	euclidean	opencv	128	205	0.56
clean	ElasticFace-Cos	cosine	ssd	120	205	0.56
clean	GhostFaceNet	cosine	opencv	54	205	0.44
clean	GhostFaceNet	euclidean	opencv	69	205	0.44
clean	GhostFaceNet	euclidean_l2	opencv	66	205	0.44

disguise:clean ethnicity:black

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
clean	ElasticFace-Cos	cosine	retinaface	111	162	1.0
clean	ElasticFace-Cos	euclidean_l2	retinaface	111	162	1.0
clean	ElasticFace-Arc+	cosine	opencv	93	162	0.88
clean	ElasticFace-Arc+	euclidean_l2	opencv	93	162	0.88
clean	ElasticFace-Arc+	cosine	retinaface	97	162	0.88
clean	ElasticFace-Arc+	euclidean_l2	retinaface	97	162	0.88
clean	ElasticFace-Cos	cosine	opencv	105	162	0.88
clean	ElasticFace-Cos	euclidean_l2	opencv	105	162	0.88
clean	QMagFace	cosine	Mtcnn	121	159	0.88
clean	QMagFace	euclidean	Mtcnn	124	159	0.88
clean	QMagFace	euclidean_l2	Mtcnn	125	159	0.88
clean	ElasticFace-Cos	cosine	ssd	107	162	0.75
clean	ElasticFace-Cos	euclidean_l2	ssd	107	162	0.75
clean	GhostFaceNet	euclidean	retinaface	95	162	0.62
clean	GhostFaceNet	euclidean_l2	retinaface	89	162	0.62
clean	ElasticFace-Arc+	euclidean	opencv	92	162	0.62
clean	ElasticFace-Arc+	euclidean_l2	ssd	89	162	0.62
clean	ElasticFace-Arc+	cosine	ssd	89	162	0.62
clean	ElasticFace-Arc+	euclidean	retinaface	91	162	0.5
clean	ElasticFace-Arc+	euclidean	ssd	89	162	0.5
clean	ElasticFace-Cos	euclidean	opencv	88	162	0.5
clean	ElasticFace-Cos	euclidean	retinaface	89	162	0.5
clean	ElasticFace-Cos	euclidean	ssd	88	162	0.5
clean	GhostFaceNet	euclidean_l2	opencv	70	162	0.5
clean	ElasticFace-Cos+	euclidean_l2	retinaface	73	162	0.38
clean	GhostFaceNet	cosine	retinaface	78	162	0.38
clean	GhostFaceNet	euclidean	opencv	75	162	0.38
clean	GhostFaceNet	euclidean	ssd	62	162	0.38
clean	ElasticFace-Cos+	cosine	retinaface	73	162	0.38
clean	ElasticFace-Cos+	cosine	opencv	71	162	0.38
clean	ElasticFace-Cos+	euclidean_l2	opencv	71	162	0.38
clean	ElasticFace-Cos+	euclidean_l2	ssd	49	162	0.25
clean	GhostFaceNet	cosine	ssd	42	162	0.25
clean	GhostFaceNet	cosine	opencv	48	162	0.25
clean	GhostFaceNet	euclidean_l2	ssd	50	162	0.25
clean	ElasticFace-Cos+	euclidean	ssd	35	162	0.25
clean	ElasticFace-Cos+	cosine	ssd	49	162	0.25
clean	ElasticFace-Cos+	euclidean	retinaface	47	162	0.25
clean	ElasticFace-Cos+	euclidean	opencv	48	162	0.25

disguise:all ethnicity:white

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
rainbowhat_off	QMagFace	euclidean_l2	Mtcnn	205	339	1.0
snood_crosses	QMagFace	euclidean	Mtcnn	216	345	0.95
halohat_off	QMagFace	euclidean_l2	Mtcnn	243	388	0.91
snood_blobs	QMagFace	euclidean_l2	Mtcnn	188	351	0.9
glasses_off	ElasticFace-Arc+	cosine	retinaface	167	375	0.86
rainbowhat_on	QMagFace	euclidean_l2	Mtcnn	203	424	0.76
snood_lines	QMagFace	euclidean_l2	Mtcnn	163	361	0.76
glasses_on	ElasticFace-Arc+	cosine	retinaface	150	365	0.73
goggles	ElasticFace-Arc+	cosine	retinaface	128	370	0.68
halohat_on	QMagFace	euclidean_l2	Mtcnn	142	428	0.45

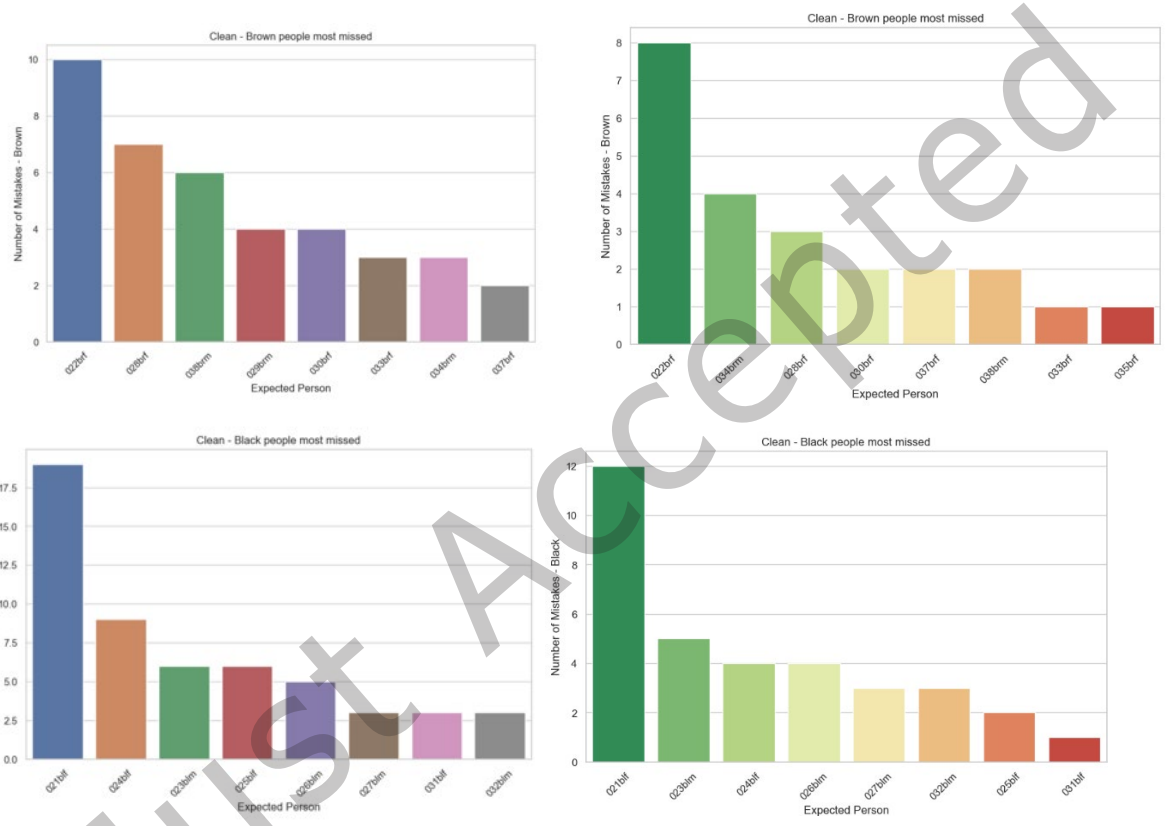
disguise:all ethnicity:brown

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
snood_lines	QMagFace	euclidean_l2	Mtcnn	163	215	1.0
snood_crosses	QMagFace	euclidean	Mtcnn	148	201	1.0
halohat_off	QMagFace	cosine	Mtcnn	145	186	1.0
halohat_on	QMagFace	euclidean_l2	Mtcnn	142	201	1.0
glasses_off	QMagFace	cosine	Mtcnn	164	196	1.0
rainbowhat_off	QMagFace	cosine	Mtcnn	143	181	1.0
rainbowhat_on	QMagFace	euclidean	Mtcnn	187	235	1.0
glasses_on	QMagFace	cosine	Mtcnn	166	211	1.0
goggles	QMagFace	euclidean_l2	Mtcnn	159	228	0.89
snood_blobs	QMagFace	euclidean_l2	Mtcnn	154	197	0.89

disguise:all ethnicity:black

Disguise	Model	Distance_Metric	Backend	Count	Total	Accuracy
glasses_off	QMagFace	euclidean_l2	Mtcnn	155	197	1.0
halohat_off	QMagFace	euclidean_l2	Mtcnn	138	176	1.0
snood_crosses	QMagFace	euclidean_l2	Mtcnn	140	186	1.0
glasses_on	ElasticFace-Cos	cosine	ssd	125	211	0.88
halohat_on	QMagFace	euclidean_l2	Mtcnn	148	200	0.88
rainbowhat_off	QMagFace	euclidean_l2	Mtcnn	130	172	0.88
rainbowhat_on	QMagFace	euclidean_l2	Mtcnn	141	191	0.88
snood_lines	ElasticFace-Arc+	cosine	retinaface	94	200	0.62
snood_blobs	ElasticFace-Arc+	cosine	retinaface	87	179	0.5
goggles	ElasticFace-Arc+	cosine	retinaface	60	200	0.38

A comparison of the *brown* participants most missed to be identified: 2023 on the left, 2024 on the right.



A comparison of the *black* participants most missed to be identified: 2023 on the left, 2024 on the right.