Investigating the Impact of Inclusion in Face Recognition Training Data on Individual Face Identification

Chris Dulhanty
chris.dulhanty@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Alexander Wong
a28wong@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

ABSTRACT

Modern face recognition systems leverage datasets containing images of hundreds of thousands of *specific* individuals' faces to train deep convolutional neural networks to learn an embedding space that maps an *arbitrary* individual's face to a vector representation of their identity. The performance of a face recognition system in face verification (1:1) and face identification (1:N) tasks is directly related to the ability of an embedding space to discriminate between identities. Recently, there has been significant public scrutiny into the source and privacy implications of large-scale face recognition training datasets such as MS-Celeb-1M and MegaFace, as many people are uncomfortable with their face being used to train dual-use technologies that can enable mass surveillance. However, the impact of an individual's inclusion in training data on a derived system's ability to recognize them has not previously been studied. In this work, we audit ArcFace, a state-of-the-art, open source face recognition system, in a large-scale face identification experiment with more than one million distractor images. We find a Rank-1 face identification accuracy of 79.71% for individuals present in the model's training data and an accuracy of 75.73% for those not present. This modest difference in accuracy demonstrates that face recognition systems using deep learning work better for individuals they are trained on, which has serious privacy implications when one considers all major open source face recognition training datasets do not obtain informed consent from individuals during their collection.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Computing methodologies** → **Visual content-based indexing and retrieval**; • **Computer systems organization** → *Neural networks*; • **Social and professional topics** → Surveillance.

KEYWORDS

face recognition, neural networks, privacy, informed consent

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375875

ACM Reference Format:
Chris Dulhanty and Alexander Wong. 2020. Investigating the Impact of Inclusion in Face Recognition Training Data on Individual Face Identification. In *2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES'20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375875

1 INTRODUCTION

Face recognition systems using Deep Convolutional Neural Networks (DCNNs) depend on the collection of large image datasets containing thousands of sets of *specific* individuals' faces for training. Using this data, DCNNs learn a set of parameters that can map an *arbitrary* individual's face to a feature representation, or *faceprint*, that has small intra-class and large inter-class variability. The ability of a face recognition system to distinguish between identities within this embedding space depends on the size and diversity of its training data, along with its model capacity and underlying algorithms. Face recognition systems have benefited from the enabling power of Internet in the collection of large-scale image datasets and from hardware improvements in enabling efficient training of large models. Recently, increased attention to face recognition by academia, industry and government has brought new researchers, ideas and funding to the field, leading to performance improvements on benchmark tasks Labelled Faces in the Wild (LFW) [20] and MegaFace [32]. Consequently, face recognition systems are now being integrated into consumer and industrial electronic devices and offered as application programming interfaces (APIs) by providers such as Amazon, Microsoft, IBM, Megvii and Kairos. However, along with improved performance has come increased public discourse on the ethics of face recognition systems and their development.

Algorithmic auditing of commercial face analysis applications has uncovered disparate performance for intersectional groups across several tasks. Poor performance for darker skinned females by commercial face analysis APIs has been reported by Buolamwini, Gebru and Raji [5, 35], as has lower accuracy in face identification by commercial systems with respect to lower (darker) skin reflectance by researchers at the US Department of Homeland Security [9]. As bias in training data begets bias in model performance, efforts to create more diverse datasets for these tasks have resulted. IBM's Diversity in Faces dataset [28], released in January 2019, is a direct response to this body of research. Using ten established coding schemes from scientific literature, researchers annotated one million face images in an effort to advance the study of fairness and accuracy in face recognition. However, this dataset has seen public scrutiny from a different, but equally notable perspective. A March 2019 investigation by NBC News into the origins of the dataset

brought to the public conversation the issue of informed consent in large-scale academic image datasets, as IBM leveraged images from Flickr with a Creative Commons Licence without notifying content owners of their use [40].

To rationalize the collection of large-scale image datasets without explicit consent of individuals, some computer vision researchers appeal to the non-commercial nature of their work. However, work by Harvey *et al.* at MegaPixels have found that authors' stated limitations on dataset use do not translate to real-world restrictions [16]. In the case of Microsoft's MS-Celeb-1M dataset, authors included an explicit "non-commercial research purpose only" clause with the dataset, which was the largest publicly-available face recognition dataset at the time. However, as the dataset has been cited in published works by the research arms of many commercial entities, findings cannot easily be isolated from improvements in product offerings. As a direct result of MegaPixel's work on the ethics, origins, and privacy implications of face recognition datasets, MS-Celeb-1M [15], Stanford's Brainwash dataset [41] and Duke's Multi-Target, Multi-Camera dataset [37] were removed from their authors' websites in June 2019. However, in the case of MS-Celeb-1M, the data remains accessible via torrents, derived datasets and other hosts [16].

In addition to issues of bias and informed consent in data collection, the general use of face recognition systems by commercial and government agencies has been raised by civil rights groups and research centers, as there is no oversight for its deployment in civil society [1, 49]. For these and other reasons, multiple cities in the United States have banned the use of face recognition systems for law enforcement purposes [8, 36, 51]. Many people are concerned with their identify being used to train the dual-use technology that is face recognition. With reports of face recognition being used by law enforcement entities to identify protesters in London [4] and Hong Kong [29], and measures enacted to ban face masks in the latter location [53], there is merit in understanding the impact of one's inclusion in the training data that fuels the development of these systems.

In an effort to inform the conversation about informed consent and privacy in the domain of face recognition, we conduct experiments on a state-of-the-art system. The goal of this work is to determine the impact of an individual's inclusion in face recognition training data on a derived system's ability to recognize them. To the best of the authors' knowledge, this is the first paper to investigate this relationship.

The remainder of this paper is organized in the following manner; section two outlines ethical considerations for some decisions in the design and implementation of this work, section three provides background for the taxonomy, algorithms and data used in face recognition research, section four outlines the design of experiments used to address the research question, section five presents our results and adds discussion and the paper concludes in section six.

2 ETHICAL CONSIDERATIONS

2.1 Intent

The intent of this work is to investigate the performance of face recognition systems with respect to inclusion in training datasets. While one interpretation of this work may be to motivate efforts to mitigate demographic bias in the development of face recognition systems, it should be noted that increasing the performance of face recognition systems in any context can increase their ability to be used for oppressive purposes. In addition, due to historical societal injustices against marginalized populations and racially-biased police practices in the United States, a disproportionate number of African Americans and Hispanics are present in mugshot databases, often used by law enforcement agencies as data sources for face recognition systems [14, 31]. These populations are therefore poised to receive a greater burden of the effects of improved face recognition systems. We therefore position this work as informing the discussion on data privacy and consent when it comes to face recognition systems and do not advocate for technical improvements without a larger discussion on the appropriate use and legality of the technology.

2.2 Use of MS-Celeb-1M

As noted in the introduction, the MS-Celeb-1M dataset was removed from Microsoft's website in June 2019. In a response to a Financial Times inquiry, Microsoft stated the website was retired "because the research challenge is over" [30]. However, a version of this dataset with detected and aligned faces from a "cleaned" subset of the original images is available from the Intelligent Behaviour and Understanding Group (iBUG) at Imperial College London. The dataset was offered as training data for the "Lightweight Face Recognition Challenge & Workshop"[1] the group organized at ICCV 2019. The group has pre-trained face recognition models available as benchmarks for the challenge, trained on this data.

As this work aims to conduct experiments in a realistic setting in order to better inform the conversation around data collection processes, the analysis of a state-of-the-art model, trained on a large dataset is necessary to gain insights that are applicable to commercial applications. We therefore have decided to use the MS-Celeb-1M dataset, through its derived version offered for the ICCV 2019 Workshop, for the limited scope of this work.

3 BACKGROUND

3.1 Face Recognition Tasks

Within the domain of face recognition lies two categories of tasks: *face verification* and *face identification* [24].

In face verification, the goal is to assess if a presented image matches with the reference image of an individual, often to grant access to a physical device or location. Unlocking a smartphone with one's face provides an example of face verification; a person presents their face to a phone and it is verified against a reference image of the known owner of the device. This task is referred to as 1:1 matching, as there is only one individual that the presented face image is compared against. In order to confirm a match, a threshold of similarity must be met, which can be set by the developer of a system to meet a specific level of security. Performance of a system on face verification tasks is reported in terms of accuracy; the number of correct verifications of all verification attempts.

[1] https://ibug.doc.ic.ac.uk/resources/lightweight-face-recognition-challenge-workshop/

Table 1: Prominent open-source face recognition training datasets

Dataset	Year Released	# Identities	# Images	Informed Consent Obtained?	Source
CASIA WebFace	2014	10,575	494K	No	[52]
CelebA	2015	10,177	203K	No	[26]
VGGFace	2015	2,622	2.6M	No	[34]
MS-Celeb-1M	2016	99,952	10.0M	No	[15]
UMDFaces	2016	8,277	368K	No	[3]
MegaFace (Challenge 2)	2016	672,057	4.7M	No	[32]
VGGFace2	2018	9,131	3.3M	No	[6]

In face identification, a *gallery* of known identities is constructed from face images of individuals in advance of testing. Subsequently, a face image of unknown identity is presented to the system as the *probe*. The probe is then matched for similarity with all images in the gallery, constituting 1:N matching. If the system guarantees that the identity of the probe is within the gallery of identities, the problem is considered *closed-set face identification*, otherwise it is considered *open-set face identification*.

Closed-set face identification tasks are common in academic benchmarks, as galleries are carefully constructed by their authors to contain all probes. In open-set face identification, a confidence threshold must be set to reject matches that do not meet a certain level of similarity. The selection of an appropriate threshold is especially relevant in high-risk applications such as law enforcement in which false positives have significant implications.

Face identification performance is reported in terms of accuracy in returning the correct identity of a probe from the gallery, or in the open-set case, no identity if the probe does not exist in the galley. Common performance metrics include Rank-1 accuracy; of all identification attempts, the number of times the correct identity in the gallery is the most similar identity to the probe, and Rank-10 accuracy; the number of times the correct identity is in the ten most similar identities to the probe.

3.2 Deep Face Recognition

Rapid improvements in image classification in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [38] by AlexNet [23], ZFNet [54], GoogLeNet [42] and ResNet [17] from 2012 to 2015 cemented the DCNN as the standard method in computer vision research and applications. While early uses of convolutional neural networks in face verification showed preliminary success [7, 19], it was not until the introduction of the aforementioned network architectures that the modern era of deep face recognition was in full swing. Coupled with innovations in loss function design and access to larger image datasets, modern face recognition systems have improved state-of-the-art performance on benchmark face verification and identification tasks significantly in the past six years. For a complete survey of the development of deep face recognition systems, please refer to the review paper by Wang and Deng [47]; the following is a brief summary of major milestones.

The first system to adapt findings from ILSVRC to face recognition was Facebook's DeepFace [43], published in 2014 by Taigman *et al.*. The nine-layer AlexNet-based model was trained on a private dataset of 4.4M images of 4K identities and achieved state-of-the-art accuracy on face verification tasks LFW and YouTube Faces (YTF) [50], reducing the error rate by more than 50% on the latter task.

Following this work, Google introduced FaceNet in 2015 with a major innovation in loss function design [39]. While the standard *softmax* loss function optimized inter-class differences, researchers found that intra-class differences remained high, problematic in the domain of face recognition. To rectify this problem, the *triplet loss* was introduced to jointly minimize the Euclidean distance between an anchor example and a positive example of the same identity and maximize the distance between an anchor and negative example. Using a ZFNet-based model and a private dataset of 200M images of 8M identities, they achieved state-of-the-art performance on LFW and YTF.

Innovations in loss functions dominated the next wave of improvements in benchmark tasks, motivated by improving discrimination between classes by making features more separable. Wen *et al.* introduced the Center Loss in 2016 [48], followed by Liu *et al.* with the Angular Softmax in 2017 [25]. The Large Margin Cosine Loss was introduced in 2018 by Wang *et al.* [46], and in 2019, Deng *et al.* incorporated the Additive Angular Margin Loss into the ArcFace model [10], considered state-of-the-art on multiple face recognition benchmarks when published.

3.3 Face Recognition Training Datasets

Access to large-scale face recognition training datasets has been essential to the development of modern solutions by the academic community. While early published resulted in the DCNN-era of face recognition came out of companies with access to massive private datasets, such as Facebook's 500M images and 10M identities [44] and Google's 200M images and 8M identities [39], the release of several open-source datasets in the ensuing years has allowed researchers to train models at scale. A summary of notable face recognition training datasets of the past six years is provided in Table 1. These datasets catalyzed the field of face recognition and lead to great advances in model performance on benchmark tasks. They largely consist of celebrity identities and copyrighted images scraped from the internet.

One exception is MegaFace, which is derived from the YFCC100M dataset of 100M photos with a Creative Commons Licence, from 550K personal Flickr accounts [45]. While the Creative Commons Licence permits the fair use of images, including in this context, Ryan Merkley, CEO of Creative Commons, noted the trouble of conflating copyright with privacy in a March 2019 statement: "... copyright is not a good tool to protect individual privacy, to address

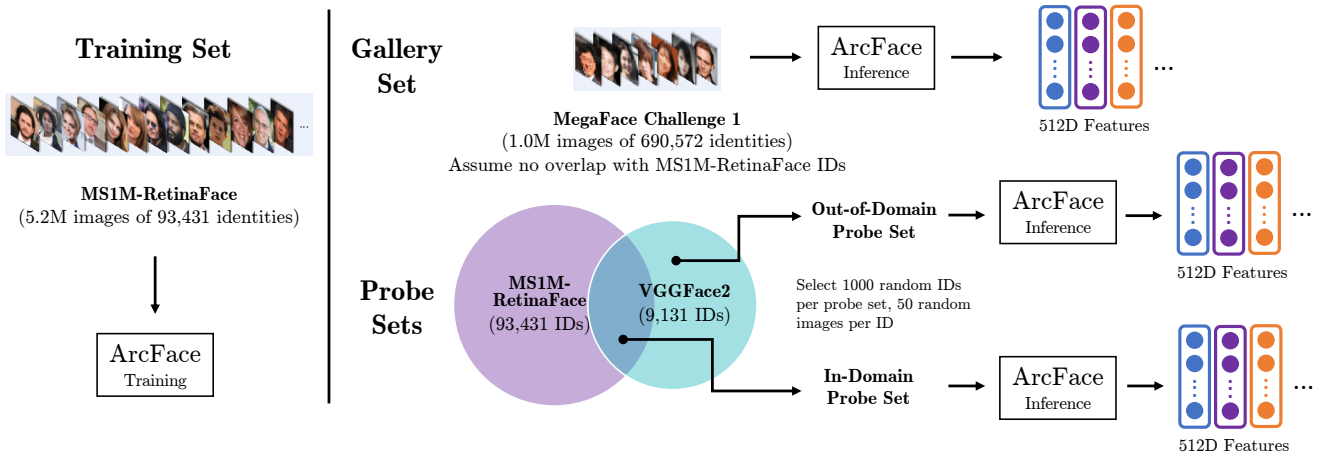

Figure 1: Experimental procedure to generate feature representations of images in gallery and probe sets from ArcFace model

research ethics in AI development, or to regulate the use of surveillance tools employed online. Those issues rightly belong in the public policy space, and good solutions will consider both the law and the community norms of CC licenses and content shared online in general" [27]. While MegaFace contains unknown, non-celebrity identities, an October 2019 investigation by the New York Times demonstrated that account metadata associated with images in the dataset allows for a trivial real-world identification of individuals [18].

In all datasets, no informed consent was sought or obtained for individuals contained therein.

4 METHODOLOGY

4.1 Face Recognition Model

4.1.1 Training Data. We employ a cleaned version of the MS-Celeb-1M dataset [15] as training data for a face recognition model in this work. This dataset was prepared for the ICCV 2019 Lightweight Face Recognition Challenge [11]. All face images were preprocessed by the RetinaFace model for face detection and alignment [12]. A similarity transformation was applied to each detected face using five predicted face landmarks to generate normalized face crops of 112 x 112 pixels.

As the original version of this dataset has been shown to exhibit considerable inter-class noise, efforts have been made to automatically clean the dataset [21]. In the case of this version, after face detection and alignment, cleaning was performed by a semi-automatic refinement strategy. First, a pre-trained ArcFace model [10] was used to automatically remove outlier images of each identity. A manual removal of incorrectly labelled images by "ethnicity-specific annotators" followed to result in a dataset of 5,179,510 images of 93,431 identities. We refer to this dataset as *MS1M-RetinaFace*.

4.1.2 Model. We select the ArcFace model [10] to study in this work. ArcFace employs the Additive Angular Margin Loss and a ResNet100 backbone to arrive at a 512-dimensional feature representation of an input image. The model achieves a verification accuracy of 99.83% on LFW and Rank-1 identification accuracy of 81.91% on the MegaFace Challenge 1 with one million distractors, considered state-of-the-art results. We select the model for study as is the top academic, open-source entrant on the National Institute of Standards and Technology (NIST) Face Recognition Vendor Test (FRVT) 1:1 Verification[2], a benchmark used by many commercial entities to validate the performance of their face recognition systems. Pre-trained weights for this model were provided by iBUG.

4.2 Experiments

To determine the effect of inclusion in the training data of a face recognition system on its ability to identify an individual, we frame the problem as a closed-set face identification task. We construct two probe datasets and perform face identification on a gallery of one million distractor images. We assess the performance of the model on the probe datasets in terms of Rank-1, Rank-10 and Rank-100 identification accuracies. A visual representation of the datasets used in this work is shown in Figure 1.

4.2.1 Probe Data. We construct two probe datasets from the VGGFace2 dataset [6]. Using regular expressions, we match identities in VGGFace2 by name with the identify list of MS1M-RetinaFace. We find 5,902 VGGFace2 identities present in MS1M-RetinaFace and 3,229 VGGFace2 identities not present in the training dataset. In each of these two groups, we randomly select 500 male identities and 500 female identities for evaluation, based on gender labels provided by VGGFace2 metadata. For each identity, we randomly select 50 images and perform face detection and alignment with the Multi-task Cascaded Convolutional Network (MTCNN) [55] to generate normalized face crops of size 112 x 112 pixels. We refer to the set of 50,000 images of 1000 identities present in the training data as the *in-domain probe set* and the set of 50,000 images of 1000 identities not present in the training data at the *out-of-domain*

[2]https://www.nist.gov/programs-projects/frvt-11-verification

Table 2: Face identification accuracies of ArcFace model on different probe image sets with one million distractor images

Metric	Probe Set	All	Males	Females
Rank-1 Accuracy (%)	In-Domain	79.71	78.50	80.93
	Out-of-Domain	75.73	77.30	74.17
Rank-10 Accuracy (%)	In-Domain	90.82	90.92	90.73
	Out-of-Domain	86.58	88.59	84.57
Rank-100 Accuracy (%)	In-Domain	92.72	92.52	92.92
	Out-of-Domain	89.22	90.59	87.84

probe set. We then generate 512-dimensional feature representations for all images in the in-domain and out-of-domain probe sets by running them through ArcFace.

4.2.2 Gallery Data. We leverage the MegaFace Challenge 1 "Distractor" dataset [22] of 1,027,058 images of 690,572 identities to form the basis of the *gallery*. We again apply MTCNN to generate normalized face crops of 112 x 112 pixels for each image and run each image through ArcFace to generate 512D feature representations of all images in the gallery.

4.2.3 Evaluation Protocol. The experiments conducted in this work follow the protocol of MegaFace Challenge 1, with our probe sets in place of the standard FaceScrub test set [33]. We employ the Linux development kit offered by MegaFace to perform evaluation. Each probe set is evaluated following Algorithm 1; a written description of this protocol follows.

A probe set contains 1000 identities, each with 50 images represented as 512D features. For each identity, we iterate over their images, adding one image to the gallery at a time, which we will refer to as *the needle*. We then iterate over the remaining 49 images, using each one as a probe. We rank all images in the gallery by L2 distance in feature space to the probe, and record the position of the needle in the ranked list. We report results for each probe set in terms of Rank-1, Rank-10 and Rank-100 face identification accuracies.

Algorithm 1: Closed-set face identification evaluation

Result: Rank-1, 10 and 100 face identification accuracies for a probe set.

$r_1, r_{10}, r_{100} = 0$;
gallery contains 1M distractor images;
for *identity in identities*$_{1\ to\ 1000}$ **do**
  for *image*$_{needle}$ *in images*$_{1\ to\ 50}$ **do**
    add *image*$_{needle}$ to the gallery;
    for *image*$_{probe}$ *in images*$_{1\ to\ 50}$ **do**
      if *image*$_{needle}$ == *image*$_{probe}$ **then**
        continue;
      else
        rank all images in gallery by L2 distance to *image*$_{probe}$ in feature space;
        if *image*$_{needle}$ *in first position in ranked list* **then**
          $r_1 = r_1 + 1$
        if *image*$_{needle}$ *in first 10 positions in ranked list* **then**
          $r_{10} = r_{10} + 1$
        if *image*$_{needle}$ *in first 100 positions in ranked list* **then**
          $r_{100} = r_{100} + 1$
    remove *image*$_{needle}$ from gallery;
Rank-$1_{\text{Acc.}} = r_1/(1000 \times 50 \times 49)$;
Rank-$10_{\text{Acc.}} = r_{10}/(1000 \times 50 \times 49)$;
Rank-$100_{\text{Acc.}} = r_{100}/(1000 \times 50 \times 49)$;

5 RESULTS AND DISCUSSION

We present results of the experiments in Table 2 for Ranks 1, 10 and 100. We find there is a modest increase in face identification accuracy for identities present in the training data, compared to those who are not. In-domain identities have a 4.0% higher identification accuracy than out-of-domain identities at Rank-1, 4.2% higher at Rank-10, and 3.5% higher at Rank-100. Although not a significant margin, these results suggest that modern DCNN-based face recognition systems are biased towards individuals they are trained on.

The disparate performance between probe sets suggests some amount of overfitting has occurred in the model. Although the model generalizes well to new identities, as evidenced by results on benchmarks LFW, MegaFace and on NIST's FRVT, these results indicate that the 93k identities the system is trained on are more easily identifiable in a large-scale study. As the model's Additive Angular Margin Loss sought to increase discrimination between classes by making features more separable, it appears the model has learned to map identities to the same feature representation more consistently for those it has seen before.

We also investigated the role of gender in the performance of the face recognition model. We find small differences in performance between genders for in-domain identities, but a 3 - 4% decrease in performance for females compared to males who are out-of-domain, across all ranks. These results suggest that a gender bias exists in the face recognition model towards female identities. As the model has a smaller drop in face identification accuracy between domains for males, it has a greater ability to generalize to new male identities. While we do not have gender labels available for all identities in MS1M-RetinaFace, recent work has demonstrated that large-scale face recognition datasets are largely biased towards lighter-skinned males [28]. A representational bias in MS1M-RetinaFace may account for this disparate performance across genders. Looking at these results in a different way, the consistent performance for

in-domain identities across genders is perhaps more evidence that the model is overfitting to identities it has seen before. If the model only had a gender bias, we would have seen disparate performance for genders on both probe sets, however, these results suggest the model may also exhibit a "training inclusion bias".

Results of this study lead to the question; is the bias towards individuals in training data truly a consequence of overtraining, or is this a fundamental element of deep face recognition models? If we look to the manner by which the model was trained, overfitting in a traditional sense seems unlikely, as early stopping was employed, and results on held-out test identities demonstrate strong generalization. Perhaps there is a generalization gap in performance between in-domain and out-of-domain identities that is not apparent in current validation protocols, and increased regularization can mitigate this gap. Further testing on different training datasets and model architectures will be necessary to gather more evidence to answer this question.

We did not analyze the effect of skin type on face recognition model performance in this study, as skin type annotations were not available to us at the time. However, two considerations were made to attempt to control for effects of skin type in these results. First, the selection of 1000 identities for each probe set is far larger than what is used in the standard protocol of MegaFace Challenge 1, where 80 identities are sampled from FaceScrub. Having a larger sample size helps to control for identities who may have either superior or poor performance due to possible model bias. In addition, the approach of random sampling in-domain and out-of-domain probe sets ensures both contain a similar distribution of identities with respect to skin type, with the assumption that the identities common to MS1M-RetinaFace and VGGFace2 and the identities distinct to VGGFace2 follow the same distribution of skin type. As both MS1M-RetinaFace and VGGFace2 use the popularity of celebrities online to construct identity lists, this assumption seems to be reasonable. Having said this, the role of skin type in the performance of the model is a very important relationship to study, and this is planned for future work. Fitzpatrick skin type [13] annotations will need to be collected for all individuals in VGGFace2 such that sampling can be done to ensure even representation in probe sets across gender and skin type, and to determine intersectional accuracy.

The results of this study are quite concerning from a privacy and informed consent perspective. As described in the background section on Face Recognition Training Datasets, there does not exist a major open-source dataset that gathers informed consent from the individuals it contains. Without these individuals' knowledge or permission, the systems trained on their identities have a greater ability to identify them. As face recognition becomes more powerful and ubiquitous, the ability for misuse becomes greater. While MS-Celeb-1M contains only "celebrity" identities, this classification of an individual should not negate informed consent in the development of powerful surveillance technologies. Face recognition systems are unique among biometrics as the face can be easily captured at distance without one's knowledge. The face uniquely identifies an individual, and it is difficult to opt-out of these systems without wearing a mask or other means of obfuscation, drawing undue attention to one's self. From a legal perspective, the concept of informed consent in the analysis of images of individuals' faces has traction in some jurisdictions. As reported by the New York Times with reference to potential financial liabilities of MegaFace [18], the Illinois Biometric Information Privacy Act [2] is a State law enacted in 2008 that gives Illinois residents the right to seek financial compensation from entities using their face scans without their informed consent.

The experiments in this work aim to simulate a real-world testing environment of a state-of-the-art face recognition system, with a gallery of more than one million images. These findings, therefore, may hold for systems that are currently deployed in the real-world.

6 CONCLUSION

In this work we present the first study to investigate the role of inclusion in face recognition training data on a derived system's ability to identify an individual. Through the construction of two sets of probe data that overlap and are distinct from the training data of a state-of-the-art system, we conduct a large-scale face identification experiment. We find a modest 4% improvement in face identification accuracy for individuals who are present in training data, which is highly problematic given the norm in the field is to not gather informed consent in the collection of training datasets. Future work will apply this methodology to more models, training datasets and distance metrics (i.e. cosine distance) to see if results are consistent. Following prior work [5, 9, 35], analysis of face recognition model bias with respect to gender, skin type and their intersections in large-scale face identification tasks is needed, as well as tying results to representational bias in training data. Additionally, the relationship between the *number of images* of an individual in training data and their ability to be identified is an interesting area of study. Finally, analysis of a face recognition model's feature space directly provides an alternative to a task-based auditing approach, and may be fruitful for understating nuances of inter- and intra-class differences.

ACKNOWLEDGMENTS

We would like to thank the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program for their support.

REFERENCES

[1] ACLU. 2018. Aclu Calls For Moratorium On Law And Immigration Enforcement Use Of Facial Recognition. https://www.aclu.org/press-releases/aclu-calls-moratorium-law-and-immigration-enforcement-use-Facial-Recognition
[2] Illinois General Assembly. 2008. 740 ILCS 14 / Biometric Information Privacy Act. http://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57
[3] Ankan Bansal, Anirudh Nanduri, Carlos D Castillo, Rajeev Ranjan, and Rama Chellappa. 2017. Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 464–473.
[4] Owen Bowcott. 2018. Police face legal action over use of facial recognition cameras. *The Guardian* (Jun 2018). https://www.theguardian.com/technology/2018/jun/14/police-face-legal-action-over-use-of-facial-recognition-cameras
[5] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.
[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 67–74.
[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Compter Vision and Pattern Recognition*. 539–546.
[8] Kate Conger, Richard Fausset, and Serge F. Kovaleski. 2019. San Francisco Bans Facial Recognition Technology. *The New York Times* (May 2019). https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html

[9] Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. 2019. Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1, 1 (2019), 32–41.

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4690–4699.

[11] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. 2019. Lightweight face recognition challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.

[12] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. *arXiv preprint arXiv:1905.00641* (2019).

[13] Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* 124, 6 (1988), 869–871.

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

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

[16] Adam Harvey and Jules LaPlace. 2019. *MegaPixels: Origins, Ethics, and Privacy Implications of Publicly Available Face Recognition Image Datasets*. https://megapixels.cc/

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Kashmir Hill and Aaron Krolik. 2019. How Photos of Your Kids Are Powering Surveillance Technology. *The New York Times* (Oct 2019). https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html

[19] Gary B Huang, Honglak Lee, and Erik Learned-Miller. 2012. Learning hierarchical representations for face verification with convolutional deep belief networks. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2518–2525.

[20] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report 07-49. University of Massachusetts, Amherst.

[21] Chi Jin, Ruochun Jin, Kai Chen, and Yong Dou. 2018. A community detection approach to cleaning extremely large face database. *Computational intelligence and neuroscience* 2018 (2018).

[22] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4873–4882.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[24] Erik Learned-Miller, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. 2016. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*. Springer, 189–248.

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 212–220.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[27] Ryan Merkley. 2019. Use and Fair Use: Statement on shared images in facial recognition AI. https://creativecommons.org/2019/03/13/statement-on-shared-images-in-facial-recognition-ai/

[28] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in faces. *arXiv preprint arXiv:1901.10436* (2019).

[29] Paul Mozur. 2019. In Hong Kong Protests, Faces Become Weapons. *The New York Times* (Jul 2019). https://www.nytimes.com/2019/07/26/technology/hong-kong-protests-facial-recognition-surveillance.html

[30] Madhumita Murgia. 2019. Microsoft quietly deletes largest public face recognition data set. *Financial Times* (Jun 2019). https://www.ft.com/content/7d3e0d6a-87a0-11e9-a028-86cea8523dc2

[31] NAACP. 2018. Criminal Justice Fact Sheet. http://www.naacp.org/criminal-justice-fact-sheet/

[32] Aaron Nech and Ira Kemelmacher-Shlizerman. 2017. Level Playing Field For Million Scale Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[33] Hong-Wei Ng and Stefan Winkler. 2014. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 343–347.

[34] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, Mark W. Jones Xianghua Xie and Gary K. L. Tam (Eds.). BMVA Press, Article 41, 12 pages. https://doi.org/10.5244/C.29.41

[35] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. 429–435.

[36] Sarah Ravani. 2019. Oakland bans use of facial recognition technology, citing bias concerns. *San Francisco Chronicle* (Jul 2019). https://www.sfchronicle.com/bayarea/article/Oakland-bans-use-of-facial-recognition-14101253.php

[37] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[39] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[40] Olivia Solon. 2019. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. *NBCNews.com* (Mar 2019). https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

[41] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. 2016. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2325–2333.

[42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[43] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[44] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2015. Web-scale training for face identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2746–2754.

[45] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. YFCC100M: The new data in multimedia research. *arXiv preprint arXiv:1503.01817* (2015).

[46] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5265–5274.

[47] Mei Wang and Weihong Deng. 2018. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655* (2018).

[48] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*. Springer, 499–515.

[49] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. AI Now Report 2018. (2018).

[50] Lior Wolf, Tal Hassner, and Itay Maoz. 2011. *Face recognition in unconstrained videos with matched background similarity*. IEEE.

[51] Sarah Wu. 2019. Somerville City Council passes facial recognition ban. *The Boston Globe* (Jun 2019). https://www.bostonglobe.com/metro/2019/06/27/somerville-city-council-passes-facial-recognition-ban/SfaqQ7mG3DGulXonBHSCYK/story.html

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

[53] Elaine Yu. 2019. Hong Kong Court Reinstates Mask Ban Before Citywide Election. *The New York Times* (Nov 2019). https://www.nytimes.com/2019/11/22/world/asia/hong-kong-mask-ban-protests-election.html

[54] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[55] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.