Concerning Trends in Facial Recognition Evaluation

Anonymous Authors

Abstract

We survey over 100 datasets dating between 1976 to 2019 collectively representing about 145 million images of over 17 million subjects from a range of sources, demographics and conditions. Our survey reveals the effect that different scientific motivations, technological advancements, and contexts have had on shaping trends in face dataset construction and evaluation. We begin with a historical survey of the field, then discuss how choices of task selection, data collection practices, and evaluation procedures have skewed our understanding of what is required for these technologies to function appropriately within their deployed context.

Introduction

Whether in schools (Shultz 2019), convenient stores (Spears 2019), public squares (Bridges 2019; Mesnik 2018; coo 2019), concerts (Bridges 2019), apartment complexes (Durkin 2019), airports (O'Flaherty 2019), neighbourhood parks (Chinoy 2019), or on a personal device (Apple Inc. 2019), facial recognition is increasingly pervading our lives in numerous, unaccountable ways. Earlier this year, the National Institute of Standards (NIST) proudly announced that between 2014 and 2018, the technology improved twenty fold, to a failure rate of just 0.2 percent (NIST 2018). Yet a string of failed real world pilots contradicts the academic mythos of facial recognition as a solved problem.

Eight trials of deployments in London between 2016 and 2018 resulted in a 96% rate of false identifications as criminal suspects(Dearden 2019). A 2019 report found that 81% of suspects flagged through the facial recognition tool used by London's Metropolitan Police were wrongly identified (Manthorpe and Martin 2019). Similarly, New York City's Metropolitan Transportation Authority (MTA) reported pilot results of a 100% error rate, meaning no one at all had been properly identified with the technology (Berger 2019). Furthermore, such failures are not evenly distributed across demographic subgroups. A study in 2018 revealed that for gender classification, commercial facial recognition API's performed up to 30% worse on a darker skinned female subgroup compared to a lighter male subgroup(Buolamwini and Gebru 2018). Its follow up 2019 audit, Actionable Auditing(Raji and Buolamwini 2019), as well as subsequent studies by NIST (Ngan and Grother) and other academics (Vangara et al. 2019) have confirmed these disparities and demonstrated they extend to other problems such as face identification and verification tasks.

Aside from functional concerns, several cities have responded to the threat of the weaponized use of the technology as a surveillance tool by banning its use completely by government actors (ACLU 2019; Yee 2019; Haskins 2019; Council of the City of Berkeley 2018; Somerville City Council 2019) and others have sought to restrict the use of the technology in certain deployment contexts (Montgomery and Hagemann 2019), such as housing (Clarke 2019) or in schools(Wallace et al. 2019). Many states have passed laws specifically addressing the privacy violations inherent in the development and operation of these systems (Hasegawa, Saldaña, and Nguyen 2019; Idaho Judiciary, Rules and Administration Committee 2019; Texas Legislature 2019; Chau 2019; Ting 2019; Cavenaugh 2019; Ritchie 2019; Castro et al. 2019; Illinois Legislature 2008), with many more states presiding over legislative proposals (Carlyle et al. 2019; Lucido 2019; Ting 2019; Creem et al. 2019; Bowers 2019; Farmer 2019) and a federal bill pending (Blunt 2019).

Yet despite the growing public awareness of the practical failure of these systems once released in the real world, academic studies continue to report near perfect performance of facial recognition systems on benchmark datasets. In an attempt to better understand this dissonance between the perceived functionality of these systems under current evaluation norms and the reality of its performance when deployed, we surveyed over 100 datasets across the history of facial recognition evaluation from its recorded beginnings in the 1960s to its present day form. This is the largest and most recent survey of this kind to date - the latest past survey was conducted in 2012 (Forczmański and Furman 2012) with only forty-one datasets; prior to that, a survey was conducted in 2005 just twenty-one datasets (Jain and Li 2011).

After identifying significant eras of facial recognition technology development, we analyze the evolution of evaluation tasks, data and metrics to gain a clearer picture of what will be required for evaluations to truly capture a reliable representation of the performance of these systems in a deployed context.

Terminology & Scope

Facial recognition technology (FRT) in this study will be considered as a broad term to encompass any task involving the identification and characterization of the face image of a human subject, including face detection – the task of locating a face within a bounding box in an image, face verification – the one-to-one confirmation of a query image to a given image, face identification – the one-to-many matching of a query image to the most similar results within a given repository of images, and facial analysis – a classification task to determine facial characteristics, including physical or demographic traits like age, gender or pose, as well as more situational traits such as facial expression.

Mainstream commercial facial recognition products are still predominantly based on still 2-D image-based predictions (Wang and Deng 2018) so we limit the scope of this survey to the consideration of 2-D still-image photographic facial recognition benchmarks that are presently available online. This omits datasets comprised of non-visual face images representing infra-red or other sensor output maps, sketch or drawing datasets, video-based datasets, 3-D image datasets and datasets addressing full body human identification.

Historical Context of Facial Recognition Development

We begin with a brief overview of the historical context of facial recognition technological development in order to anchor our understanding of the major shifts that defined the evolution and technical progress of this technology, which in turn shapes the themes of concurrent shifts in the norms of evaluation for this technology.

Era I: Early Research Findings (1964 - 1995)

In 1964, Woodrow Bledsoe first attempted "facial recognition" in a computational form. Funded by an "undisclosed intelligence agency" and armed with a book of mugshots and a probe photograph, he used a computer program to connect the identity of the suspect to an identity in the book of mugshots (Bledsoe 1966). While Bledsoe's approach of matching vectors of facial landmark features became popular, it was computationally expensive and slow; with the technology at the time, Bledsoe could only process around 40 pictures an hour (Bledsoe 1966). Eventually, a new method called eigenfaces, which represented the pixel intensity of face features in a lower dimensional space, offered an appealing alternative approach. However, obtaining enough data at the time to attempt such new methods was challenging, as researchers had to recruit and hire models and photographers, manually design the set up for consistent or controlled illumination, and manually label data, including facial landmarks (Jafri and Arabnia 2009).

Era II: Commercial Viability as the "New Biometric" (1996 - 2006)

By 1996, government officials had recognized and embraced the face as a non-invasive biometric attribute that could be used to track and identify individuals without requiring their explicit physical participation (Phillips et al. 2000). The Face Recognition Technology (FERET) dataset was thus created with $6.5 million of funding from the U.S. Department of Defense and the National Institute of Standards and Technology (NIST) to provide researchers the data they required to make progress in the field. In 15 photography sessions of the same set up between August 1993 and July 1996, images were collected in a semi-controlled environment (P.J. Phillips and Rauss 2000). The resulting benchmark began with 2,413 still face images, representing 856 individuals, and grew to contain 14,126 facial images of 1,199 individuals, available upon request. At the moment of its release, it became the largest and most comprehensive effort to create a benchmark that would accurately compare and evaluate existing facial recognition algorithms(P.J. Phillips and Rauss 2000). The large data effort, as well as the corresponding government sponsored facial recognition algorithm development competitions and research investments (Phillips et al. 2005), proved to be successful in igniting academic research interest in the field. By the end of the program, the field had matured from an under explored research problem to spawning a growing industry of commercial products.

Attempts at commercialization with these early methods revealed that even small environmental changes, such as in image illumination and a subject's pose, could at this time be enough to obscure or distort the features required to make a match. Similarly, any unexpected change in their face - from aging to a new facial expression to partial occlusions, such as a scarf, mask, or pair of glasses – could cripple the performance of the technology(Sharif et al. 2017; Forczmański and Furman 2012; Yang, Kriegman, and Ahuja 2002).

Era III: Mainstream Development for Unconstrained Settings (2007-2013)

Labeled Faces in the Wild (LfW) filled a need for researchers to access more naturally situated and varied data. The dataset leveraged the Web to source the first fully unconstrained face dataset with over 13,000 images of 1,680 faces in an infinite combination of poses, illumination conditions, and expressions (Huang et al. 2007).

The Labeled Faces in the Wild (LfW) dataset inspired a flurry of Web-scraped face datasets for facial recognition model training and benchmarking - including many datasets sourcing images without consent from online platforms such as Google Image search (Bainbridge, Isola, and Oliva 2013; Han et al. 2017; Cao et al. 2018b), Youtube (Chen et al. 2017; Dantcheva, Chen, and Ross 2012), Flickr (Merler et al. 2019; Kemelmacher-Shlizerman et al. 2016) and Yahoo News (Jain and Learned-Miller 2010). As the appetite for unstructured and unconstrained "in the wild" data grew, there was also in this period a proliferation of benchmarks like ChokePoint(Wong et al. 2011) and SCface (Grgic, Delac, and Grgic 2011), datasets that source face images from mock or real surveillance set ups.

The research problem of identifying faces in unconstrained conditions nevertheless remained a stubborn technical challenge and development stalled as academics strug-

gled to develop methods to represent faces independently of a controlled image context and template appearance.

Era IV: Deep Learning Breakthrough (2014 and onwards)

It was not until the breakthrough of Alexnet in 2012, and the subsequent introduction of the DeepFace model in 2014, that the use of neural networks became a mainstream method for Facial recognition development. DeepFace was the first instance of a facial recognition model approaching human performance on a task. Deepface was developed by researchers at Facebook, Inc. and trained on an internal dataset composed of images from Facebook profile images; at the time, it was purportedly "the largest facial dataset to-date, an identity labeled dataset of four million facial images belonging to more than 4,000 identities" (Taigman et al. 2014).The impact of deep learning techniques on face recognition and its adjacent problems was dramatic; the DeepFace model achieved a 97.35% accuracy on the Labeled Faces in the Wild (LfW) test set, reducing the previous state of the art's error by 27%.

In response to this technological advance, the size of subsequently constructed face datasets grew significantly to accommodate the growing data requirements to train deep learning models. The rapid progress sparked high commercial interest, as well. Moving beyond security applications, facial recognition products began to encompass use cases that include "indexing and searching digital image repositories", "customized ad precise delivery", "user engagement monitoring" and "customer demographic analysis"(Phillips et al. 2005).

Survey of Facial Recognition Evaluation

We execute a historical survey of 133 datasets created between 1976 to 2019. The datasets collectively representing 145,143,610 images of 17,733,157 individual people's faces. Celeb 500k of 2018 is the largest dataset, containing 50,000,000 images (Cao, Li, and Zhang 2018), and the FRVT Ongoing challenge data from NIST contains the most image subjects, including the faces of 14,400,000 (Grother, Ngan, and Hanaoka 2018). The smallest dataset is 54 images of 4 people from 1988's JACFEE (Japanese and Caucasian Facial Expressions of Emotion) dataset (Biehl et al. 1997). Overall, on average here are 1,262,118 images and 159,758 subjects.

We then do a chronological analysis of these currently accessible face datasets. We note trends in the design decisions made with the release of these benchmarks and datasets and map how such trends feed into or result in current misunderstandings of the limitations of this technology upon deployment. The full details of the datasets included in the survey can be found linked here.

Task Selection

Tasks are highly influenced by who is creating and funding the dataset. At times, especially for government datasets, the goal of the developed technology is explicit and specifically defined in the design of the evaluation - for instance, the NIST FRVT dataset is funded by the Department of Homeland Security and contains data sourced from "U.S. Department of State's Mexican non-immigrant Visa archive"(Phillips et al. 2003). The prioritized and dominant use case for this technology is thus still security, access control, suspect identification, and video surveillance in the context of law enforcement and security (Sharif et al. 2017; Zhao et al. 2003), as we can see from the historical context that the government promoted and supported this technology from the start for the purpose of enabling criminal investigation and surveillance.

More diverse applications, such as the integration into mobile devices, robots, and smart home facility User Interfaces (Wang and Deng 2018), monitoring user engagement or social objectives such as finding missing children (NIST 2015) emerge only in Era IV. Facial analysis tasks emerge only in the most recent era as well. The exceptions to this are emotion datasets, which, with much older benchmarks, are often datasets sourced from the Psychology field and repurposed as evaluations of machine learning models.

Facial analysis is the class of tasks that is likely to include the most ambiguous model objectives, often implicating the "discredited pseudosciences of physiognomy and phrenology" (Metz 2019), where a subject's inner state is wrongly inferred through the evaluation of that subject's external features. Pseudo-scientific tasks to predict " sexual orientation" (Wang and Kosinski ; Leuner 2019), "attractiveness" (Eisenthal, Dror, and Ruppin 2006; Schmid, Marx, and Samal 2008), "hireability" (Fetscherin, Tantleff-Dunn, and Klumb 2019), "criminality" (Wu and Zhang 2016), and even more accepted but contested attributes, such as affect (Picard 2000), gender(Keyes 2018) and race (Benthall and Haynes 2019), are rarely questioned in the evaluation of the system, and the potential for use cases to cause harm not often considered during system testing.

Following the introduction of Amazon's Mechanical Turk service in 2005 (mtu 2019), researchers began attempting to clean and make sense of their data, while also enabling the datasets to be used to address additional tasks. Certain data and meta-labels sourced for the images are controversial. For instance, the CelebA dataset contains five landmark locations, and forty binary attributes annotations per image. These labels include the problematic and potentially insulting labels regarding size - "chubby", "double chin" - or inappropriate racial characteristics such as "Pale skin" "Pointy nose", "Narrow eyes" for Asian subjects and "Big nose" and "Big lips" for many Black subjects. Additionally there is the bizarre inclusion of concepts, such as "bags under eyes", " 5 o'clock shadow" and objectively impossible labels to consistently define, such as "attractive" (Liu et al. 2015).

Benchmark Data

Face data is biometric information as unique and identifiable as a fingerprint, but is also passively collected and thus incredibly vulnerable to perpetuating severe privacy violations. The history of which tasks we chose to explore and which issues the community was attempting to address has an important impact on the evolution on the nature of face data benchmarks developed in this field.

Table 1: Historical Arcs of Facial Recognition Development.

Era	Era 1 (before 1996)	Era 2 (1996 - 2007)	Era 3 (2007-2014)	Era 4 (2014+)
Number of Datasets Created	5	37	33	45
Range of number of Images in a dataset (MIN- MAX)	56 - 14,126	120 - 121,589	154 - 750,000	642 - 50,000,000
Range of number of Subjects in a dataset (MIN- MAX)	4 - 1,199	10 - 37,437	32 - 40,395	50 - 14,400,000
Average number of images in a dataset	2,032	11,250	46,308	2,620,489
Average number of subjects in a dataset	136	1,641	4,078	75,726

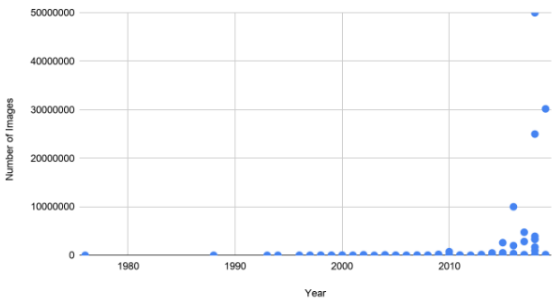

Figure 1: Summary of face dataset size over the analyzed time period of 1976 to 2019. FRVT Ongoing challenge data from NIST contains the most image subjects, including the faces of 14,400,000.

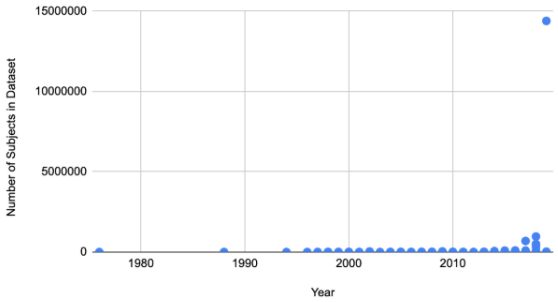

Figure 2: Subject population size over observed period of 1976 to 2019.Celeb 500k of 2018 is the largest dataset, containing 50,000,000 images.

Dataset Size Following the demonstration of the effectiveness of deep learning on facial recognition tasks following the release of DeepFace (Taigman et al. 2014), there became a growing awareness of the need for larger scale datasets. Datasets grew from tens of thousands of images to millions in likes of MegaFace and VGG-Face2. The goal was to create datasets large enough dataset to avoid overfitting and have enough of a variance to be meaningful, yet also of acceptably data quality (Wang and Deng 2018). One can also aim to set up a benchmark with depth, which has a limited number of subjects but many images for each subjects (such as VGGFace2 (Cao et al. 2018a)) or a dataset breadth, meaning the set contains many subjects but limited images for each subject (such as MS-Celeb-1M (Guo et al. 2016) and Megaface (Kemelmacher-Shlizerman et al. 2016).

Data Sources Early on, when data requirements for model development were low, the common practice was to set up *photo shoots* in order to capture face data controlled for pose, illumination, and expression. Subject consent for participation and data distribution as well as photo ownership are often mentioned explicitly in references for datasets with photography data sources. Depending on the scale of these projects, producing high quality datasets in this vein was highly expensive. And for such a set up, details like camera equipment specifications would matter in determining the quality of the image and overall dataset.

As an alternative, datasets were also sometimes a *collection* curated from other image datasets perhaps built for a different purpose, or simply *crowdsourced* from willing participants who donated their face data after being convinced or paid to do so. Many government collection sources for face data include specifically *mugshots*, often of "deceased persons with prior multiple encounters" (Founds et al. 2011). In addition to this, stills from webcam footage and official documentation such as VISA photos (Ngan, Ngan, and Grother 2015).

Later academic and corporate sources tended to derive more from the Web (Kemelmacher-Shlizerman et al. 2016; Parkhi et al. 2015; Huang et al. 2007; Guo et al. 2016) through *web searches* for still-image examples of "unconstrained" faces, or by taking frames from online videos. Some databases also tapped *surveillance camera* footage to mine face data (Grgic, Delac, and Grgic 2011; Ristani et al. 2016; Stewart, Andriluka, and Ng 2016). In these cases, the cameras were a set up in a cafe, school campus or public square (Ristani et al. 2016; Stewart, Andriluka, and Ng 2016)- effectively a more subversive photo shoot to capture "in the wild" data. Either case can often be seen as a violation of subject consent.

The diversification of data sources from photography sessions to more crowdsourced and Web-based data sources allowed for a greater diversity of subjects and image conditions, all at a much lower cost than previous attempts. However, in exchange for more realistic and diverse datasets, there was also a loss of control, as it became unmanageable to obtain subject consent, record demographic distributions, maintain dataset quality and standardize attributes such as image resolution across Internet-sourced datasets.

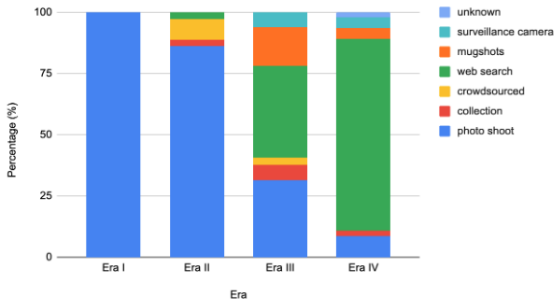

Figure 3: Data source.

Data Sharing & Reporting Datasets constructed from photo shoots in Era I and II pay considerable attention to issues of copyright and the protection of image ownership rights, with papers for benchmark datasets from these periods often indicating the informed consent of individuals participating in a photoshoot (P.J. Phillips and Rauss 2000). For example, the report describing the FRVT 2000 challenge benchmark dataset comments: "The subjects appearing in the images are all unpaid volunteers who had been briefed on the purpose of their participation and who had positively consented to the study. For privacy reasons the data was gathered anonymously" (Blackburn, Bone, and Phillips 2001).

Distribution of these datasets from this period was at times physically restricted, as even academic datasets required sending images via physical hard drives, incurring a cost to distribution that disappeared with the shift to online options (Biehl et al. 1997; **?**). Once datasets became accessible online, consent and privacy became difficult to manage. Certain strategies, such as sharing hyperlinks rather than the downloaded images, deanonymizing identities, restricting online dataset access or aiming to crawl the photos of celebrities and public figures only emerged in later eras to address this challenge. The level of awareness of privacy concerns seemed to differ greatly - at times, unconsenting adult subjects were included in datasets available for direct online download (Bainbridge 2012), and other times, distribution of this biometric information was handled sensitively, completely closed off to the public and evaluated exclusively through a custom API or graphic user interface, such as NIST's Biometric Experimentation Environment (BEE) test environment. However, many situations are somewhere in between both extremes, with dataset access granted following a formal request and agreement to the presented terms of use. Data disclosure and distribution practice can also be culturally specific. For instance, the Iranian dataset (Bastanfard, Nik, and Dehshibi 2007) includes female subjects but specifies not to allow for the public display and distribution of this female subgroup specifically, likely due to cultural restrictions of their exposure.

Dataset Reporting The reporting of datasets is wildly unstandardized. Many datasets lack information about the source and methodology by which images are collected, or fail to include information at the macro (e.g. demographic) and micro (e.g. image specific attributes or metadata creating) level, producing an incomplete picture of the dataset characteristics. Datasets might be described in an academic paper and/or on a project website, with no standardized format of disclosure, and potential inconsistencies even across different communication mediums and references. For instance, in several cases, the number of images reported on a website might differ from the number in the published paper - and at times both numbers could contradict the size reported in a survey paper or subsequent study working with the dataset (Forczmański and Furman 2012). This indicates a lack of provenance and reporting norms to track and appropriately communicate about face dataset construction and evolution.

Interestingly, some of the most comprehensive reporting was performed by NIST as part of their series of FRVT challenges, which are ongoing. Evaluation reports meticulously document the construction (source and method of collections) of their benchmark data. They acknowledge the importance of doing so in their 2000 evaluation report: "Image collection and archival are two of the most important aspects of any evaluation. Unfortunately, they do not normally receive enough attention during the planning stages of an evaluation and are rarely mentioned in evaluation reports." (Blackburn, Bone, and Phillips 2001)

Demographic Representation Although a recently revived topic of interest, researchers flag the propensity for racial bias in the FRVT dataset (Blackburn, Bone, and Phillips 2001), and even indicate model performance disparities over gender and age as early as 2002 (Phillips et al. 2003). Understanding the existence of the issue, a surprising number of datasets, especially in Era I and Era II report the limitations of the demographic distributions of the presented dataset, with some even choosing to focus the entire dataset on one demographic, such as Asian-Celeb (Li et al. 2017), Iranian Faces (Bastanfard, Nik, and Dehshibi 2007) and Indian Faces(Lazarus, Gupta, and Panda 2018). Online sourced datasets seemed to shift towards a Western media default for demographic representation, and, as the datasets were so large, the phenomenon was difficult to track and had been until recently largely unreported (Merler et al. 2019). Several datasets have been built in response to recent awareness of this issue in order to specifically address the dearth of diversity in mainstream facial analysis benchmarks (Wang et al. 2018; Chen, Chen, and Hsu 2015; Merler et al. 2019; Buolamwini and Gebru 2018).

In some cases, the hunt for increased demographic diversity may result in inappropriate privacy violations. This is most evident with the LfW+ dataset (Han et al. 2017), where Google Image search results for keywords such as "baby", "kid", and "teenager" were used to identify juvenile images to supplement to mainly adult subjects of Labeled Faces in the Wild (LfW) (Huang et al. 2007). This dataset and others - from NIST's CHEXIA dataset (Flanagan 2015), to CAFE, a child affective dataset (LoBue and Thrasher 2015) - rarely involve even the parental consent of involved parties, putting juveniles at risk by exposing their sensitive biometric infor-

mation.

Evaluation Criteria

The way in which the evaluation process of a model occurs embeds certain insights as to what makes a particular approach to evaluation reliable and more widely influential in defining the norms of evaluation practice in facial recognition.

Consistency of Results In order for an audit to be reliable, there needs to be a guaranteed consistency to the benchmarks being used - both in terms of ethical expectations and standards, as well as the data itself. Data consistency can become especially difficult with the introduction of Web-sourced data, as urls become obsolete. Inconsistencies in ethical expectations and performance standards can also make comparison difficult from year-to-year. One element of process that is yet to be standardized is auditing scheduling - currently there is no timing mentioned as a key component of audit procedure, and without the anticipation of a regular audit period then there is no expectation for regular compliance with expectations.

Updates to equipment such as digital cameras can affect benchmark attributes such as data resolution. Within our survey, the range of photo sizes and resolutions across benchmarks is large - from 32x32 to 30722048 or even larger. As the number of pixels constitutes the direct input to methods such as deep learning, it becomes difficult to understand which element of reported performance metrics are dependent on these other variables.

Metrics There are effectively two groups of evaluations - that of a biometric evaluation for facial recognition and face identification tasks as well as that of classification accuracy for facial analysis tasks. The biometric matching process resembles image similarity search and ranking as a task, and metrics are anchored to a binary output of a match or no match. Meanwhile, classification is really about the assignment of a test example to a class category that matches the pre-determined ground truth label.

For biometric evaluation, the outcome is binary. Given two predictive outcomes - negative (ie. no match) or positive (ie. a match), we designate N to be all negatively predicted outcomes and P to be all positively predicted outcomes. If a negative prediction is true, it becomes a "True negative"(TN) result, otherwise we can designate it a "False negative" (FN) result. Similarly, if a positive result is correct, it becomes a "True Positive" (TP), counter to a "False Positive" (FN) if such is not the case. False Match Rates (FMR), and False Non-Match Rates(FNMR) are the primary metrics used for facial recognition evaluation, and are at times reported across a range of decision thresholds.

Community Adoption Another thing to consider is the level of community adoption of a particular data benchmark and its influence on facial recognition development. Collectively, the analyzed face datasets are known to be cited at least 74,211 times - implying an incredibly wide reach.

Qualitative Assessments There is an opportunity to include holistic evaluations of the product and fold that into a larger audit process. The FVRT developed by NIST, for instance, was a two-part audit process involving a "Recognition Performance Test", which was a quantitative assessment of accuracy, and the "Product Usability Test", involving a more qualitative evaluation of the ease in making use of the system in deployment (Ngan, Ngan, and Grother 2015).

Table 2: Most Influential Face Datasets per Era.

Era	Dataset Name	Citations	Year Created
Era I	Picture of Facial Affect	5,163	1976
Era II	FERET	8,126	1996
Era III	Labeled Faces in the Wild	3,746	2007
Era IV	VGGFace	2,547	2015

Recommendations

Facial recognition evaluation has evolved rapidly over the last few decades, and we are just beginning to understand how these changes must impact our understanding of the performance of a facial recognition system upon deployment. The current level of documentation seems insufficiently comprehensive, suggesting the need for data reporting standards to be created, particularly given the amount of inconsistency in data reporting. The era of dataset creation through image retrieval at scale on the web raises serious concerns around privacy, ownership, and consent–a legal question that computer scientists should also be actively engaging. In addition, some of the most critical details about a facial recognition system, such as its context of deployment, it's technical limitations and appropriate scope of use, are missing and/or not communicated within the evaluation process in anyway. A more contextual evaluation is necessary to address and communicate all the risks of this technology and determine if it should be released in society at its current scale of deployment, and this should arguably be considered in the creation and distribution of a dataset

At minimum, an important intervention moving forward is to standardize documentation practice, of the model and the face datasets meant to be used in development or evaluation. Proposals such as Model Cards (Mitchell et al. 2018), Datasheets for Datasets (Gebru et al. 2018), or the Data Privacy Label (Kelley et al. 2009) can provide reporting guidelines to support the development of such practice as the new normal in the field.

Conclusion

Facial recognition technologies pose complex ethical and technical challenges. Neglecting to unpack this complexity - to measure it, analyze it and then articulate it to others - is a disservice to those, including ourselves, who are most impacted by its careless deployment. Dataset evaluation is a critical juncture at which we can provide transparency and even accountability over facial recognition systems, and interrogate the ethics of a given dataset towards producing more responsible machine learning development.

References

[ACLU 2019] ACLU. 2019. Community Control Over Police Surveillance.

[Apple Inc. 2019] Apple Inc. 2019. About face id advanced technology.

[Bainbridge, Isola, and Oliva 2013] Bainbridge, W. A.; Isola, P.; and Oliva, A. 2013. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General* 142(4):1323.

[Bainbridge 2012] Bainbridge, W. 2012. 10k us adult faces database.

[Bastanfard, Nik, and Dehshibi 2007] Bastanfard, A.; Nik, M. A.; and Dehshibi, M. M. 2007. Iranian face database with age, pose and expression. *Machine Vision* 50–55.

[Benthall and Haynes 2019] Benthall, S., and Haynes, B. D. 2019. Racial Categories in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 289–298. New York, NY, USA: ACM. event-place: Atlanta, GA, USA.

[Berger 2019] Berger, P. 2019. Mta's initial foray into facial recognition at high speed is a bust.

[Biehl et al. 1997] Biehl, M.; Matsumoto, D.; Ekman, P.; Hearn, V.; Heider, K.; Kudoh, T.; and Ton, V. 1997. Matsumoto and ekman's japanese and caucasian facial expressions of emotion (jacfee): Reliability data and cross-national differences. *Journal of Nonverbal behavior* 21(1):3–21.

[Blackburn, Bone, and Phillips 2001] Blackburn, D. M.; Bone, M.; and Phillips, P. J. 2001. Face recognition vendor test 2000: evaluation report. Technical report, DEFENSE ADVANCED RESEARCH PROJECTS AGENCY ARLINGTON VA.

[Bledsoe 1966] Bledsoe, W. W. 1966. The model method in facial recognition. *Panoramic Research Inc., Palo Alto, CA, Rep. PR1* 15(47):2.

[Blunt 2019] Blunt, R. 2019. S.847 - 116th Congress (2019-2020): Commercial Facial Recognition Privacy Act of 2019.

[Bowers 2019] Bowers, R. 2019. Arizona HB2478 | 2019 | Fifty-fourth Legislature 1st Regular.

[Bridges 2019] Bridges, E. 2019. Facial recognition tech is creeping into our lives – i'm going to court to stop it.

[Buolamwini and Gebru 2018] Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT)*.

[Cao et al. 2018a] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018a. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

[Cao et al. 2018b] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018b. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 67–74. IEEE.

[Cao, Li, and Zhang 2018] Cao, J.; Li, Y.; and Zhang, Z. 2018. Celeb-500k: A large training dataset for face recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2406–2410. IEEE.

[Carlyle et al. 2019] Carlyle; Palumbo; Wellman; Mullet; Pedersen; Billig; Hunt; Liias; Rolfes; Saldaña; Hasegawa; and Keiser. 2019. Washington State Legislature.

[Castro et al. 2019] Castro, C.; Holmes, L.; Bush, M.; Collins, J. Y.; Peters, R.; and Murphy, L. M. 2019. Illinois General Assembly - Bill Status for SB1719.

[Cavenaugh 2019] Cavenaugh, F. 2019. Arkansas HB1943 | 2019 | 92nd General Assembly.

[Chau 2019] Chau. 2019. Bill Text - AB-1281 Privacy: facial recognition technology: disclosure.

[Chen et al. 2017] Chen, C.; Dantcheva, A.; Swearingen, T.; and Ross, A. 2017. Spoofing faces using makeup: An investigative study. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 1–8. IEEE.

[Chen, Chen, and Hsu 2015] Chen, B.-C.; Chen, C.-S.; and Hsu, W. H. 2015. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia* 17:804–815.

[Chinoy 2019] Chinoy, S. 2019. We built an 'unbelievable' (but legal) facial recognition machine.

[Clarke 2019] Clarke, Y. D. 2019. Text - H.R.4008 - 116th Congress (2019-2020): No Biometric Barriers to Housing Act of 2019.

[coo 2019] 2019. How to improve your crowd control strategy with smart crowd monitoring.

[Council of the City of Berkeley 2018] Council of the City of Berkeley. 2018. Surveillance Technology Use and Community Safety Ordinance.

[Creem et al. 2019] Creem, C. S.; Lewis, J. P.; Robinson, M. D.; and Stanley, T. M. 2019. Bill S.1385.

[Dantcheva, Chen, and Ross 2012] Dantcheva, A.; Chen, C.; and Ross, A. 2012. Can facial cosmetics affect the matching accuracy of face recognition systems? In *2012 IEEE Fifth international conference on biometrics: theory, applications and systems (BTAS)*, 391–398. IEEE.

[Dearden 2019] Dearden, L. 2019. Facial recognition wrongly identifies public criminals 96% of time, figures reveal.

[Durkin 2019] Durkin, E. 2019. New york tenants fight as landlords embrace facial recognition cameras.

[Eisenthal, Dror, and Ruppin 2006] Eisenthal, Y.; Dror, G.; and Ruppin, E. 2006. Facial attractiveness: beauty and the machine. *Neural Computation* 18(1):119–142.

[Farmer 2019] Farmer, G. 2019. FL - S1270.

[Fetscherin, Tantleff-Dunn, and Klumb 2019] Fetscherin, M.; Tantleff-Dunn, S.; and Klumb, A. 2019. Effects of facial features and styling elements on perceptions of competence, warmth, and hireability of male professionals. *The Journal of Social Psychology* 0(0):1–14.

[Flanagan 2015] Flanagan, P. A. 2015. Chexia face recognition.

[Forczmański and Furman 2012] Forczmański, P., and Furman, M. 2012. Comparative analysis of benchmark datasets for face recognition algorithms verification. In *International Conference on Computer Vision and Graphics*, 354–362. Springer.

[Founds et al. 2011] Founds, A. P.; Orlans, N.; Genevieve, W.; and Watson, C. I. 2011. Nist special databse 32-multiple encounter dataset ii (meds-ii). Technical report.

[Gebru et al. 2018] Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumeé III, H.; and Crawford, K. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

[Grgic, Delac, and Grgic 2011] Grgic, M.; Delac, K.; and Grgic, S. 2011. Scface–surveillance cameras face database. *Multimedia tools and applications* 51(3):863–879.

[Grother, Ngan, and Hanaoka 2018] Grother, P. J.; Ngan, M. L.; and Hanaoka, K. K. 2018. Ongoing face recognition vendor test (frvt) part 2: Identification. Technical report.

[Guo et al. 2016] Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 87–102. Springer.

[Han et al. 2017] Han, H.; Jain, A. K.; Wang, F.; Shan, S.; and Chen, X. 2017. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence* 40(11):2597–2609.

[Hasegawa, Saldaña, and Nguyen 2019] Hasegawa; Saldaña; and Nguyen. 2019. Washington State Legislature.

[Haskins 2019] Haskins, C. 2019. Oakland Becomes Third U.S. City to Ban Facial Recognition.

[Huang et al. 2007] Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

[Idaho Judiciary, Rules and Administration Committee 2019] Idaho Judiciary, Rules and Administration Committee. 2019. Idaho H0118 | 2019 | Regular Session.

[Illinois Legislature 2008] Illinois Legislature. 2008. 740 ILCS 14/ Biometric Information Privacy Act.

[Jafri and Arabnia 2009] Jafri, R., and Arabnia, H. R. 2009. A survey of face recognition techniques. *Jips* 5(2):41–68.

[Jain and Learned-Miller 2010] Jain, V., and Learned-Miller, E. 2010. Fddb: A benchmark for face detection in unconstrained settings.

[Jain and Li 2011] Jain, A. K., and Li, S. Z. 2011. *Handbook of face recognition*. Springer.

[Kelley et al. 2009] Kelley, P. G.; Bresee, J.; Cranor, L. F.; and Reeder, R. W. 2009. A nutrition label for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, 4. ACM.

[Kemelmacher-Shlizerman et al. 2016] Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 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.

[Keyes 2018] Keyes, O. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):1–22.

[Lazarus, Gupta, and Panda 2018] Lazarus, M. Z.; Gupta, S.; and Panda, N. 2018. An indian facial database highlighting the spectacle problems.

[Leuner 2019] Leuner, J. 2019. A Replication Study: Machine Learning Models Are Capable of Predicting Sexual Orientation From Facial Images. *arXiv:1902.10739 [cs]*. arXiv: 1902.10739.

[Li et al. 2017] Li, D.; Zhang, X.; Song, L.; and Zhao, Y. 2017. Multiple-step model training for face recognition. In *International Conference on Applications and Techniques in Cyber Security and Intelligence*, 146–153. Springer.

[Liu et al. 2015] Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[LoBue and Thrasher 2015] LoBue, V., and Thrasher, C. 2015. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in psychology* 5:1532.

[Lucido 2019] Lucido, P. J. 2019. Michigan SB0342 | 2019-2020 | 100th Legislature.

[Manthorpe and Martin 2019] Manthorpe, R., and Martin, A. J. 2019. 81% of 'suspects' flagged by met's police facial recognition technology innocent, independent report says.

[Merler et al. 2019] Merler, M.; Ratha, N.; Feris, R. S.; and Smith, J. R. 2019. Diversity in Faces. *arXiv preprints* arXiv:1901.10436.

[Mesnik 2018] Mesnik, B. 2018. How face recognition works in a crowd.

[Metz 2019] Metz, C. 2019. Facial Recognition Tech Is Growing Stronger, Thanks to Your Face. *The New York Times*.

[Mitchell et al. 2018] Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2018. Model cards for model reporting. *CoRR* abs/1810.03993.

[Montgomery and Hagemann 2019] Montgomery, C., and Hagemann, R. 2019. Precision Regulation and Facial Recognition. 4.

[mtu 2019] 2019. Amazon Mechanical Turk. Page Version ID: 925398917.

[Ngan and Grother] Ngan, M., and Grother, P. Face recognition vendor test (frvt) performance of automated gender classification algorithms.

[Ngan, Ngan, and Grother 2015] Ngan, M.; Ngan, M.; and Grother, P. 2015. Face recognition vendor test (FRVT) performance of automated gender classification algorithms. Government technical report, US Department of Commerce, National Institute of Standards and Technology.

[NIST 2015] NIST. 2015. Chexia Face Recognition.

[NIST 2018] NIST. 2018. Nist evaluation shows advance in face recognition software's capabilities.

[O'Flaherty 2019] O'Flaherty, K. 2019. Facial recognition at u.s. airports. should you be concerned?

[Parkhi et al. 2015] Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; et al. 2015. Deep face recognition. In *bmvc*, volume 1, 6.

[Phillips et al. 2000] Phillips, P. J.; Martin, A.; Wilson, C. L.; and Przybocki, M. 2000. An introduction evaluating biometric systems. *Computer* 33(2):56–63.

[Phillips et al. 2003] Phillips, P. J.; Grother, P. J.; Micheals, R. J.; Blackburn, D. M.; Tabassi, E.; and Bone, M. 2003. Face recognition vendor test 2002: Evaluation report. Technical report.

[Phillips et al. 2005] Phillips, P. J.; Flynn, P. J.; Scruggs, T.; Bowyer, K. W.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; and Worek, W. 2005. Overview of the face recognition grand challenge. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, 947–954. IEEE.

[Picard 2000] Picard, R. W. 2000. *Affective Computing*. MIT Press. Google-Books-ID: GaVncRTcb1gC.

[P.J. Phillips and Rauss 2000] P.J. Phillips, Hyeonjoon Moon, S. R., and Rauss, P. 2000. The feret evaluation methodology for face-recognition algorithms. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22, 1090 – 1104. IEEE.

[Raji and Buolamwini 2019] Raji, I. D., and Buolamwini, J. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI

products. In *Prof. of the Conference on Artificial Intelligence, Ethics, and Society*.

[Ristani et al. 2016] Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, 17–35. Springer.

[Ritchie 2019] Ritchie. 2019. New York State Assembly | Bill Search and Legislative Information.

[Schmid, Marx, and Samal 2008] Schmid, K.; Marx, D.; and Samal, A. 2008. Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. *Pattern Recognition* 41(8):2710–2717.

[Sharif et al. 2017] Sharif, M.; Naz, F.; Yasmin, M.; Shahid, M. A.; and Rehman, A. 2017. Face recognition: A survey. *Journal of Engineering Science & Technology Review* 10(2).

[Shultz 2019] Shultz, J. 2019. Spying on children won't keep them safe.

[Somerville City Council 2019] Somerville City Council. 2019. Ordinance No. 2019-16 | Code of Ordinances | Somerville, MA | Municode Library.

[Spears 2019] Spears, M. 2019. 'look at camera for entry': Tacoma convenience store using facial recognition technology.

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

[Taigman et al. 2014] Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 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.

[Texas Legislature 2019] Texas Legislature. 2019. Business and Commerce Code Chapter 503. Biometric Identifiers.

[Ting 2019] Ting. 2019. Bill Text - AB-1215 Law enforcement: facial recognition and other biometric surveillance.

[Vangara et al. 2019] Vangara, K.; King, M. C.; Albiero, V.; Bowyer, K.; et al. 2019. Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

[Wallace et al. 2019] Wallace; Epstein; Mosley; MG, M.; Simon; Gottfried; L, R.; Reyes; Otis; Simotas; Quart; Kim; Rodriguez; Fahy; Abinanti; and Weprin. 2019. New York State Assembly | Bill Search and Legislative Information.

[Wang and Deng 2018] Wang, M., and Deng, W. 2018. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*.

[Wang and Kosinski] Wang, Yilun Wang, M. K., and Kosinski, M. Deep neural networks can detect sexual orientation from face.

[Wang et al. 2018] Wang, M.; Deng, W.; Hu, J.; Peng, J.; Tao, X.; and Huang, Y. 2018. Racial faces in-the-wild: Reducing racial bias by deep unsupervised domain adaptation. *CoRR* abs/1812.00194.

[Wong et al. 2011] Wong, Y.; Chen, S.; Mau, S.; Sanderson, C.; and Lovell, B. C. 2011. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *CVPR 2011 WORKSHOPS*, 74–81. IEEE.

[Wu and Zhang 2016] Wu, X., and Zhang, X. 2016. Automated Inference on Criminality using Face Images. *ArXiv* abs/1611.04135. arXiv: 1611.04135.

[Yang, Kriegman, and Ahuja 2002] Yang, M.-H.; Kriegman, D. J.; and Ahuja, N. 2002. Detecting faces in images: A survey. *IEEE Transactions on pattern analysis and machine intelligence* 24(1):34–58.

[Yee 2019] Yee, Walton, R. H. 2019. Acquisition of surveillance technology.

[Zhao et al. 2003] Zhao, W.; Chellappa, R.; Phillips, P. J.; and Rosenfeld, A. 2003. Face recognition: A literature survey. *ACM computing surveys (CSUR)* 35(4):399–458.