Researchers Gone Wild

Origins and Endpoints of Image Training Datasets Created "In the Wild"

Adam Harvey and Jules LaPlace

Abstract

Face recognition and biometric research are contributing to rapid growth in new biometric surveillance technologies. But many of the datasets used for these technologies rely on media collected from non-consensual, nonregulated sources. Researchers refer to this media as being "in the wild." This analysis examines the widespread and largely unregulated use of images "in the wild" that were captured from campuses, CCTV camera feeds, social media, celebrity databases, and by scraping internet search engines. The findings presented here show that millions of individuals have unknowingly been used for training face recognition and other biometric analysis algorithms in both academic and commercial applications. Data compiled for this project, along with more in-depth analyses for each dataset, is available on the research project website https://megapixels.cc [https://exposing.ai/].

Overview

Image training datasets are an essential technical component of artificial intelligence (AI) that often operate out of sight. Without sufficiently large datasets to train on, AI could not compute anything. As Chris Darby, president and CEO of In-Q-Tel (C.I.A.'s strategic investment firm) has stated: "[A]n algorithm without data is useless" (Darby 2019). Geoffrey Hinton describes the importance of datasets as central to understanding new forms of computation. "Our relationship to computers has changed" Hinton says. "Instead of programming them, we now show them and they figure it out" (Hinton 2017). Kai-Fu Lee claims that "AI is basically run on data" (Lee 2019). The new logic is not better algorithms; it is better data, and more data. "The more data the better the AI works, more brilliantly than how the researcher is working on the problem" says Lee. But if data is the bedrock of AI systems, the foundation should be solid. Instead, many of the datasets currently used to train, test and validate face recognition and other biometric surveillance technologies are built on an unstable foundation of data collected without consent or oversight. Researchers call this approach "in the wild," referring to the assumed

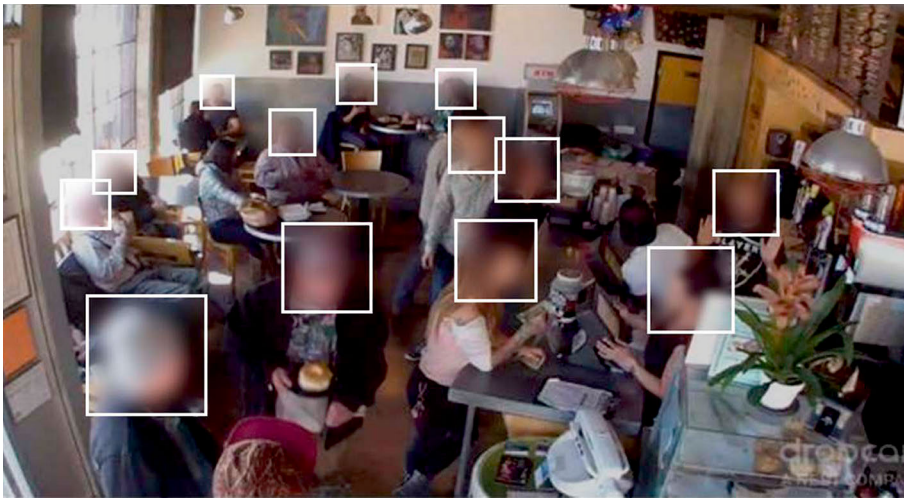

Fig. 1 A still frame from the Brainwash dataset created by researchers at Stanford University that was later discovered in a research paper linked to a foreign military organization.

natural or unconstrained quality of the data. This research examines where these datasets originate and where they are being used. In response to this research, several dataset authors have retracted or stopped distributing their datasets, deleted websites or issued formal apologies for ethical breaches in their data collection methods. A list of retracted datasets with more information is available on the MegaPixels project website.

Background

Within the field of AI, face recognition is one of the most concerning applications. In the United States, over a dozen cities have now banned face recognition, citing civil rights concerns and the potential for abuse by law enforcement agencies. Face recognition technologies are also disproportionately more threatening because of the decreased accuracy for minority racial groups (Burton-Harris and Mayor 2020; Grother, Ngan, and Hanaoka 2019), reflecting the biased data they were trained with. Several researchers have addressed the former issue by creating more diverse face datasets, and companies have responded by pledging to address this bias with algorithms that learn more from underrepresented classes. But another issue remains: How are these datasets being created, and is anyone consenting to being used for biometric research experiments?

This answer requires a collective voice, as face recognition algorithms are a collective technology that requires millions of faces from millions of people. Determining the similarity of one person to another requires the encoded knowledge of multiple identities. A face recognition system's utility is its capacity to understand the difference between a theoretically limitless variety of biometric appearances. But this assumes a limitless pool of training data and a complementary scale of computational power. In reality, resources are limited. Academic researchers seeking to participate in the field of face recognition or face analysis have long sought open, shareable resources to innovate new ideas. In the United States, where a significant amount of academic facial recognition research was funded as a response to the 9/11 attacks, universities lacked access to datasets, which were critical to advancing research. According to the authors of FERET, the first public face recognition benchmark, "[t]wo of the most critical requirements in support of producing reliable

face-recognition systems are a large database of facial images and a testing procedure to evaluate systems" (Phillips et al. 1997).

In 2007, a landmark face recognition benchmarking dataset called Labeled Faces in The Wild (LFW) (Huang et al. 2008) was first introduced to address these requirements. LFW is based on a previous dataset collected in 2003 called "Names and Faces in the News" that contained half a million captioned news images from Yahoo! News (Berg et al. 2004). The LFW dataset includes 13,233 images from 5,749 individuals. According to an article on BiometricUpdate.com, a popular site for biometric industry professional, LFW eventually became "the most widely used evaluation set in the field of facial recognition" (Lee 2017). The success of the LFW face dataset helped catalyze and normalize the trend for "media in the wild." Over the next decade, researchers replicated their success in dozens more datasets.

At the same time, corporations were amassing far larger face datasets, but these were off limits to academic research and public benchmarks. Google reportedly built an internal dataset of over 200 million images and 8 million identities, while Facebook has over 500 million images from 10 million identities (Bansal et al. 2017). Engineers and researchers who were locked out of the proprietary data sources controlled by corporations or government agencies sought alternative data collection methods to make face recognition research more widely accessible. This led to the surge in the usage and development of biometric datasets created "in the wild" that could be used for both training and public benchmarking.

Methodology

To understand how this shift in data collection has evolved, our research developed a system to categorize, track and visualize datasets by analyzing and geocoding the research citations associated with each dataset. Rather than rely on the researcher's initial intended purpose and purview to understand how a dataset is impacting society, our analysis reframes datasets as a biometric commodity in a global information supply chain. Our research maps the transnational flow of datasets that power a growing crisis of biometric surveillance technologies.

The geographic information for each dataset is inferred by using the author's stated affiliation in the front matter of publicly available research papers. Typically, a research paper is published coinciding with the release of a new dataset. This becomes the starting point to then analyze the geographic metadata in other research publications that cite the original work. For each dataset, hundreds or thousands of PDFs are located and then manually reviewed to verify whether the researchers would have needed to download the data in order to conduct their research. Specifically, we look for researchers using datasets as part of a research methodology, including as training, fine-tuning or verification data. Research papers are omitted that only mention the dataset in passing as related research, or which purely cite methodology in the original paper such as algorithms or pre-trained models, because this does not prove that the researchers acquired the images. Because our geocoding method makes an

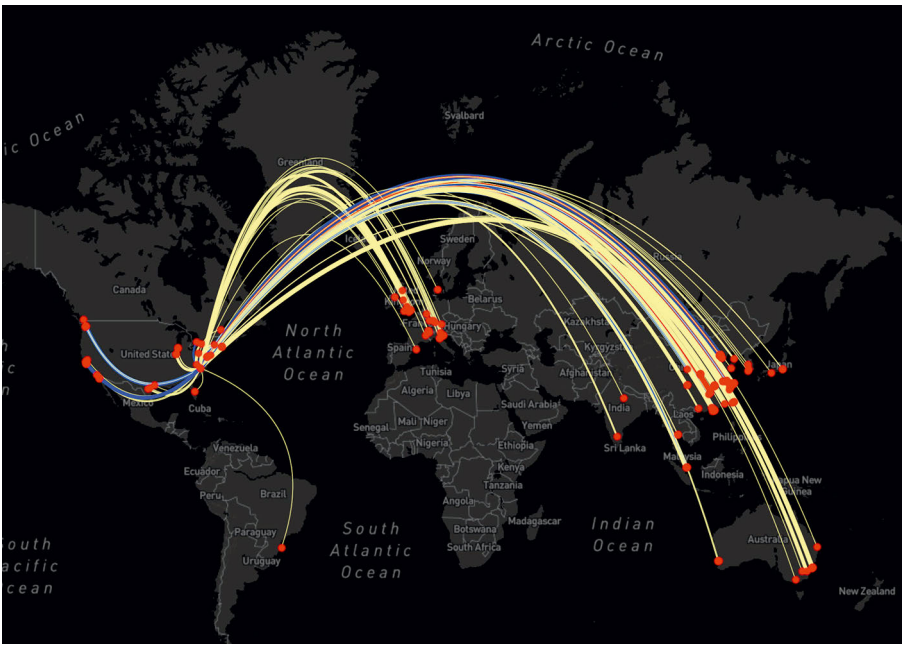

Fig. 2 A visualization of the inferred usage locations for the Duke MTMC datasets. The data originated at university campus in North Carolina and eventually became one of the most widely cited training datasets for building surveillance technologies.

assumption that each researcher has self-reported the correct organization, and that they were operating at that location during the time of their research, the inferred geographic data should be understood as a proxy for revealing the global trends associated with a particular dataset, with each point on the map understood as a confirmed usage but in an approximate location.

Next, the verified set of research papers are again reviewed for insights into the types of organizations using the data. As an example, we applied this methodology to a dataset created from a cafe in San Francisco called Brainwash (Stewart, Andriluka, and Ng 2016). The dataset was created by a researcher at Stanford University who recorded the cafe's livestream with the help of AngelCam.com and then used it to create a head detection dataset. Publicly available research uncovered during our analysis showed that images from the cafe in San Francisco were eventually used by the National University of Defense Technology in China, a military research organization affiliated with the People's Liberation Army. As a result of our investigation, the dataset was terminated by Stanford University, which triggered mainstream media articles cautioning against this type of rogue data acquisition (Metz 2019). Brainwash is one example among hundreds which are being analyzed and published on our research project website megapixels.cc that follow a similar trend.

Datasets "in the wild"

Among the hundreds of other datasets created in the wild, over a dozen inherit the same nomenclature of the popular LFW dataset. An illustrative but non-exhaustive list of datasets using "in the wild" in their title include "300 Faces In-the-Wild," "Affect in the Wild Challenge," "Annotated Faces in the Wild," "Annotated Facial Landmarks in the Wild," "Appearance-based Gaze Estimation in-the-Wild," "Biased Faces in the Wild," "Caltech Occluded Faces in the Wild," "Cross Pose Labeled Faces in the Wild," "Disguised Faces in the Wild," "Expression in the Wild," "Faces in the Wild," "Families in the Wild," "Grouping Face in the Wild (GFW) Dataset," "High Quality Faces in the Wild," "HUST-LEBW Eyeblink in the Wild Dataset," "ibug Deformable Models of Ears In-the-Wild," "In-the-Wild Child Celebrity," "Labeled Face Parts in the Wild," "LAOFIW – Labeled Ancestral Faces in the Wild," "Makeup in the Wild," "Person Re-Identification in the Wild," "Racial Faces in the Wild" and "Valence and Arousal Estimation In-the-Wild."

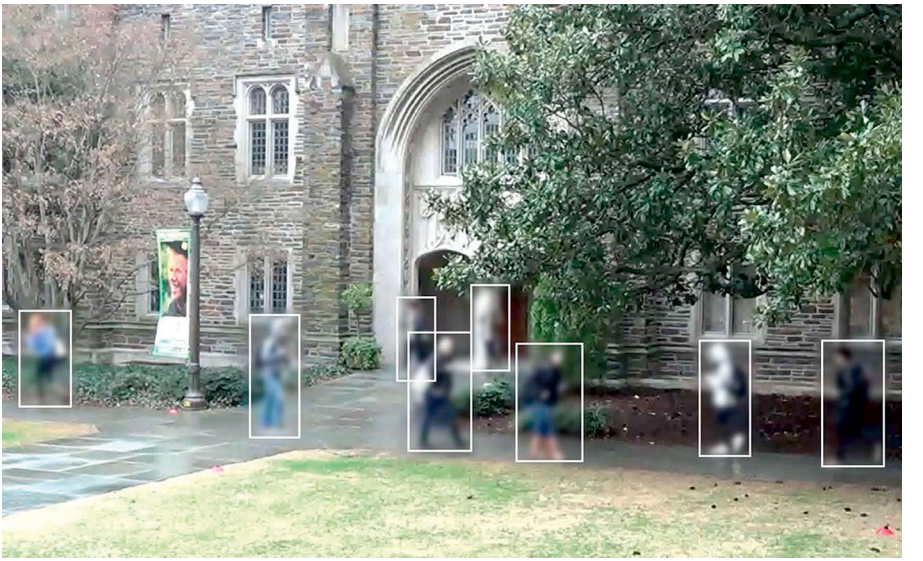

Fig. 3 A collage of images from Duke MTMC dataset recorded at Duke University without consent from the students.

Datasets make reference to their predecessors, using similar format for the files and annotations. But every dataset is unique and custom-made. There is no standard way data is collected or represented in a training dataset. And because dataset usage changes over time, an existing dataset might transform into a new dataset by being edited, re-annotated or combined with other sources. Instead of grouping the datasets by their initial intended applications or formats, which are unstable over time, we use their origin as a taxonomy for classification, broadly grouping the datasets into four themes: campuses and universities, CCTV or livestream feeds, social media, celebrity databases and web scraping.

Campus images of students

Images of students collected on campuses appear frequently in unconstrained datasets "in the wild." In the United States, several datasets were discovered that exploited campuses as a source of training data.

Duke MTMC

In 2016, a researcher at Duke University in North Carolina created a dataset of student images called Duke MTMC, or multi-target multi-camera. The Duke MTMC dataset contains over 14 hours of synchronized surveillance video from eight cameras at 1080p and 60 FPS, with over 2 million frames of 2,000 students walking to and from classes. The eight surveillance cameras deployed on campus were specifically set up to capture students "during periods between lectures, when pedestrian traffic is heavy" (Ristani et al. 2016). The dataset became widely popular and over 100 publicly available research papers were discovered that used the dataset. These papers were analyzed according to methodology described earlier to understand the endpoints: who is using the dataset and how it is being used. The results show that the Duke MTMC dataset spread far beyond its origins and intentions in academic research projects at Duke University. Since its publication in 2016, more than twice as many research citations originated in China as in the United States. Among these citations were papers linked to the Chinese military and several companies known to provide Chinese authorities with the oppressive surveillance technology used to monitor millions of Uighur Muslims.

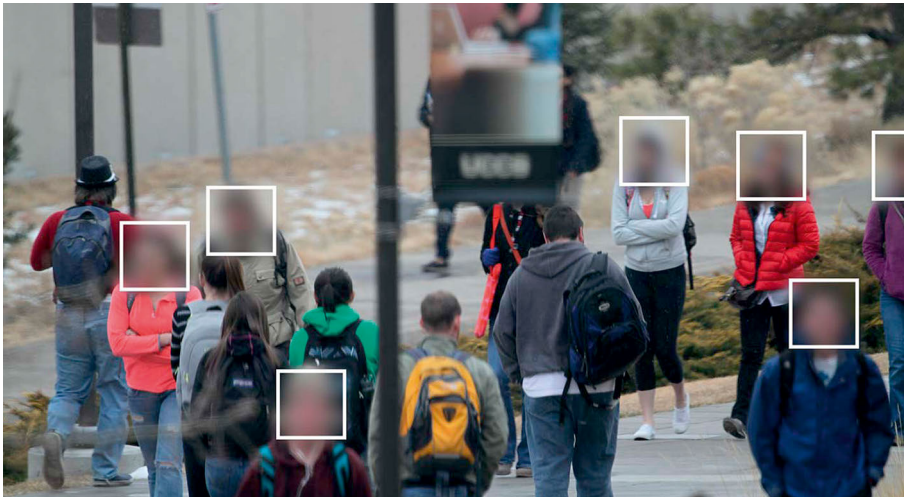

Fig. 4 A collage of images from UCCS dataset where students were photographed with a long-range camera without their awareness and used for facial recognition experiments.

In one 2018 research paper jointly published by researchers from SenseNets and SenseTime entitled "Attention-Aware Compositional Network for Person Reidentification" (Xu et al. 2018), the Duke MTMC dataset was used for "extensive experiments" on improving person re-identification across multiple surveillance cameras, with important applications in suspect tracking. Both SenseNets and SenseTime have provided surveillance technology to monitor Uighur Muslims in China (Mozur 2019).

Despite warnings that the authoritarian surveillance used in China represents a humanitarian crisis (Watch 2019), researchers at Duke University unknowingly continued to provide open access to their dataset for anyone to use for any project. As the surveillance crisis in China grew, so did the number of citations with links to organizations complicit in the crisis. In 2018 alone, there were over 90 research projects happening in China that publicly acknowledged using the Duke MTMC dataset. Among these were projects from CloudWalk, Hikvision, Megvii (Face++), SenseNets, SenseTime, Beihang University, China's National University of Defense Technology and the PLA's Army Engineering University, several of which have been added to a trade blacklist by the United States Commerce Department.

In response to our research and joint investigation with the Financial Times (Murgia 2019), the author of Duke MTMC terminated their website. The local student newspaper then published several articles about the issue, and the author responded with a formal apology to the student body, admitting that the dataset was a violation of Duke's ethics standards.

The UnConstrained College Students dataset

A similar story occurred at a campus in Colorado where university faculty used a long-range high-resolution surveillance camera and photographed students without their knowledge for a face recognition benchmarking dataset called UnConstrained College Students (UCCS).

The UCCS dataset includes over 1,700 unique identities of students and faculty walking to and from class. The photos were taken during the spring semesters of school year 2012–2013 on the West Lawn of the University of Colorado, Colorado Springs campus, using a Canon 7D 18-megapixel digital camera fitted with a Sigma 800mm F5.6 EX APO DG HSM telephoto lens, pointed out an office window

across the university's West Lawn. "The camera [was] programmed to start capturing images at specific time intervals between classes to maximize the number of faces being captured" (Günther et al. 2017). Their setup made it impossible for students to know they were being photographed, providing the researchers with realistic, unconstrained, surveillance images to help build face recognition systems for real world applications by defense, intelligence and commercial partners. In fact, the dataset was funded by the Intelligence Advanced Research Projects Activity (IARPA), the Office of Director of National Intelligence (ODNI), Office of Naval Research and the Department of Defense Multidisciplinary University Research Initiative (ONR MURI), and the Special Operations Command and Small Business Innovation Research (SOCOM SBIR). A University of Colorado, Colorado Springs website also explicitly states that their involvement in the IARPA Janus face recognition project has been developed to serve the needs of national intelligence, establishing that the dataset of student images was created in the interest of United States defense and intelligence agencies.

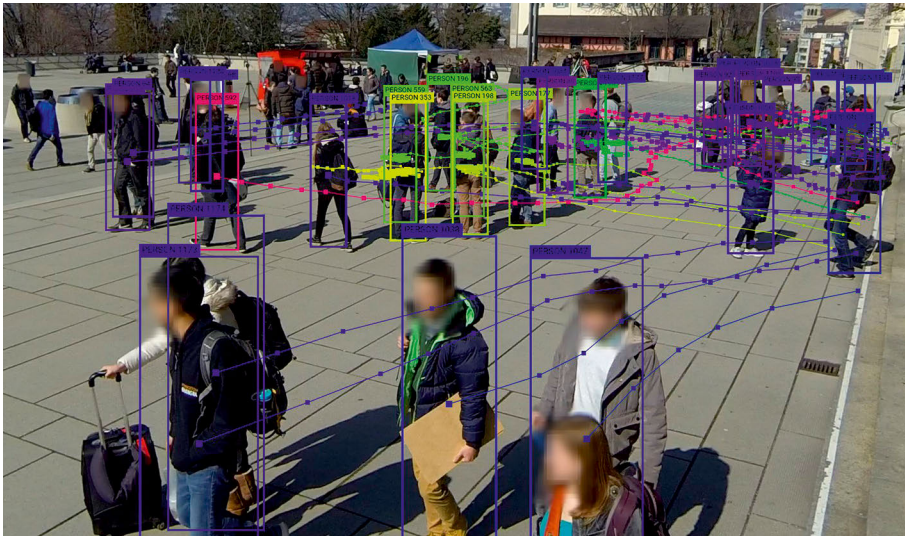

Fig. 5 A still image from Wildtrack dataset collected at ETH Zurich, where researchers recorded students and publicly distributed their videos for surveillance research. The image is annotated to track students across multiple video frames.

Wildtrack dataset

In another dataset originating in Zurich, Switzerland called Wildtrack, researchers made video recordings of students outside the ETH university main building. The videos were acquired in an "unscripted," "non-actor but realistic environment" (Chavdarova et al. 2017), implying forced consent. In total, seven 35-minute videos containing thousands of students were surreptitiously recorded and made publicly available for any type of research. Though the researchers described posting signs to inform students of what was happening (Kormann 2020), reviewing the videos shows that the vast majority of students were nonplussed. One student gave a camera the middle finger and then walked away.

The dataset eventually surfaced in a research paper on unmanned aerial vehicle (UAV) surveillance at the International Conference on Systems and Informatics, where researchers affiliated with Nanjing University of Aeronautics and the University of Leicester proposed a new method for detecting and tracking small targets from UAV surveillance feeds with applications for "conducting aerial surveillance" (Xiang et al. 2019). Figures published in their research paper confirm that video recordings of students at ETH Zurich were used for research and development of foreign UAV surveillance technologies.

CCTV and livecam images

On October 27, November 13 and November 24 in 2014, a researcher at Stanford worked with Angelcam.com (Stewart, Andriluka, and Ng 2016) to create a dataset called Brainwash. The dataset includes 11,917 images of "everyday life of a busy downtown cafe" captured at 100 second intervals throughout the day. The Brainwash dataset is notable, as mentioned previously, because the images of people in a San Francisco cafe were eventually used in multiple projects by researchers affiliated with the National University of Defense Technology in China. Brainwash is no longer distributed by Stanford, but unlike researchers at Duke, the researchers at Stanford did not provide any apology or admission of ethical breach.

Images from CCTV or security cameras provide another frequent source of data. In this case, the data is most similar to the potential environment in which it would be deployed, but the scale of these datasets is often smaller and less accessible. Datasets created

from CCTV feeds include MrSub and Clifton, datasets of surveillance images from a sandwich shop used for head detection; Grand Central Station Dataset, CCTV videos from Grand Central Station in New York City used for pedestrian tracking; QMUL GRID, a dataset of commuters from the London Underground that was released by the UK Ministry of Defence for the development of person tracking technologies; and Oxford Town Centre, a dataset of pedestrians in Oxford originally created for the development of head stabilization technologies used in face recognition systems.

The Oxford Town Centre CCTV video was obtained from a surveillance camera at the corner of Cornmarket and Market St. in Oxford, England and includes approximately 2,200 people. Since its publication in 2009 (Benfold and Reid 2011), the Oxford Town Centre dataset has been used in over 60 verified research projects including research affiliated with Amazon, Disney, OSRAM, Sony, Volvo and Huawei; and academic research in China, Israel, Russia, Singapore, the US and Germany, among dozens more.

Fig. 6 Still images from the Brainwash dataset created from a livecam feed from a cafe in San Francisco used in multiple research projects for developing head detection algorithms.

The Oxford Town Centre dataset is unique in that it uses footage from a public surveillance camera that would otherwise be designated for public safety. The video shows that the pedestrians act normally and unrehearsed indicating they neither knew of nor consented to participation in the research project. In June 2020, the website for Oxford Town Centre was taken down with no announcement or apology from the researchers.

Social media images

Social media images provide the second largest source of data "in the wild," with Flickr.com as the single largest source of data for face recognition and face analysis related experiments. The largest dataset, though not entirely comprised of faces, is called "Yahoo! Flickr Creative Commons 100 Million" or YFCC100M. As the name implies, it includes 100 million media objects with Creative Commons licenses. The YFCC100M dataset is the origin of one of the largest publicly available face recognition training datasets, called MegaFace.

MegaFace (Nech and Kemelmacher-Shlizerman 2017) is a large-scale, public face recognition training dataset that serves as one of the most important benchmarks for commercial face recognition vendors. It includes 4,753,320 faces of 672,057 identities from 3,311,471 photos downloaded from 48,383 Flickr users' photo albums. All photos included a Creative Commons license, but most were not licensed for commercial use.

MegaFace has appeared in research projects affiliated with Alibaba, Amazon, Google, CyberLink, IntelliVision, N-TechLab (FindFace.pro), Mitsubishi, Orion Star Technology, Philips, Samsung, SenseTime, Sogou, Tencent and Vision Semantics, to name only a few. A public records request by New York Times reporter Kashmir Hill revealed that the dataset has also been used by the Turkish Police, Danish National Police, Russian security and defense contractor Stilsoft, American defense contractor Northrop Grumman and Hoan Ton-That, the founder of controversial face recognition company Clearview.ai. Additionally, according to the press release from the University of Washington where the dataset was created, "more than 300 research groups [were] working with MegaFace" as of 2016, most of which are commercial. A New York Times investigation into the MegaFace dataset located and interviewed several people whose

photos were in the dataset, most of whom were disturbed to learn how their photos were being used (Hill 2019).

Images from Flickr were also used to build the Who Goes There and GeoFaces datasets, which were used for racial and ethnicity profiling in research projects that tried to convert a face into a GPS location (Islam, Workman, and Jacobs 2015).

Other datasets exploiting Creative Commons for facial training data include People In Photo Albums (PIPA), a dataset created by researchers from Facebook to improve face recognition algorithms; Labeled Ancestral Faces in The Wild (LAOFIW), which used Flickr images for ethnicity profiling; Adience, a dataset of Flickr images used for age and gender estimation algorithms; IBM Diversity in Faces, a dataset of images derived from the YFCC100M Flickr dataset and used to address bias in commercial face recognition research; and Flickr Faces High Quality (FFHQ), another dataset of Flickr images created by researchers from NVIDIA and used for synthetic face generation experiments.

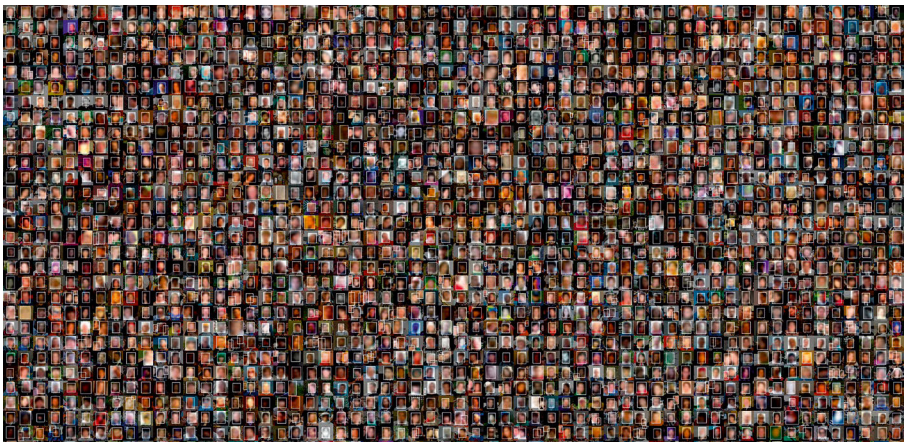

Fig. 7 Example images from the MegaFace dataset.

In total, our research discovered over 30 datasets using Flickr images. Many of these datasets overlap or comprise combinations of other datasets. Not all are used explicitly for face recognition, though all datasets have, in different ways, contributed the growth of remote biometric surveillance and analysis technologies. For example, images

from the Microsoft Common Objects in Context (MS-COCO) dataset are used for person and object detection, and person detection overlaps with person re-identification surveillance technologies. Images from the USED and RESEED datasets were primarily used for social event or activity recognition, and activity priority objective in the Defense Advanced Research Projects Activity (DARPA) Mind's Eye program (Bouma et al. 2012). Taxonomies often overlap and datasets become reused over time, but often this can be traced back to defense or military applications. For example, the LFW dataset was originally created by academics at the University of the Massachusetts and it later received funding from the Central Intelligence Agency and the National Security Agency (Jain, Learned-Miller, and McCallum 2007).

Collectively, the datasets we analyze can be described as contributing to remote biometric analysis, with overlapping applications in hard biometrics (face recognition), soft biometrics (gender, age and facial attributes), social relationship analysis (interface analysis within groups), person re-identification and activity recognition, which collectively align with the advancement of surveillance technologies in in commercial and defense applications.

Public figure and celebrity images

The largest source of data "in the wild" is images of celebrities and public figures. Though this data is less "wild" because it comprises publicity and event photos with a cast of celebrities that often reflects structural inequalities in a society and replicates their bias, it also provides a higher quantity of images per person, which enables new types of face research. Because public figures and celebrities can remain popular over time, datasets have been created to exploit individual age diversity over decades of photos.

For example, the Cross-Age Celebrity Dataset (Chen, Chen, and Hsu 2015) uses photos from 2,000 subjects in the Internet Movie Database (IMDb) to construct a facial recognition training dataset capable of recognizing people with age disparities from a query face photo. IMDb is cited as the source for several more celebrity face recognition datasets, including CASIA-Webface (Yi et al. 2014), a dataset of 10,575 subjects; and IMDb-Wiki (Rothe, Timofte, and Gool 2015), a dataset of 20,284 subjects used mostly for age and gender estimation research.

The largest source of face recognition training data is the Microsoft Celeb (MS-Celeb-1M) dataset (Guo et al. 2016). It includes 10,000,000 images from 100,000 subjects, with a target list of 900,000 more subjects, bringing the total list of names used in the project to 1,000,000. Microsoft's goal in building this dataset was to distribute an initial training dataset of 100,000 individuals' biometric data to accelerate research into recognizing a larger target list of one million people "using all the possibly collected face images of [these] individual[s] on the web as training data" (Guo et al. 2016).

While the majority of people in this dataset are American and British actors, the exploitative use of the term "celebrity" extends far beyond Hollywood. Many of the names in the MS-Celeb face recognition dataset are merely people who must maintain an online presence for their professional lives: journalists, artists, musicians, activists, policy makers, writers and academics. Many people in the target list are even vocal critics of the very technology Microsoft is using their name and biometric information to build. It includes digital rights activists like Jillian York; artists critical of surveillance including Trevor Paglen, Jill Magid and Aram Bartholl; Intercept founders Laura Poitras, Jeremy Scahill and Glenn Greenwald; Data and Society founder danah boyd; Shoshana Zuboff, author of "Surveillance Capitalism"; and even Julie Brill, the former FTC commissioner responsible for protecting consumer privacy.

Microsoft didn't only create MS-Celeb for other researchers to use; they also use it internally. In a publicly available 2017 Microsoft Research project called "One-shot Face Recognition by Promoting Underrepresented Classes," Microsoft used the MS-Celeb face dataset to build their algorithms and advertise the results. Microsoft's corporate version of the paper does not mention that they used the MS-Celeb dataset, but the open-access version published on arxiv.org does, stating that Microsoft analyzed their algorithms "on the MS-Celeb-1M low-shot learning benchmark task" (Guo and Zhang 2017).

Despite the recent termination of the msceleb.org website, the dataset still exists in several repositories on GitHub and on the hard drives of countless researchers, on AcademicTorrents.org, and will likely continue to be used in research projects around the world. For example, the MS-Celeb dataset was used for a competition called "Lightweight Face Recognition Challenge & Workshop," where the

best face recognition entries received monetary awards. The organizers of the workshop provide the MS-Celeb-1M data as a 250GB file containing the cropped faces (iBug 2019).

In June 2019, after Microsoft had taken down the dataset website, MS-Celeb reemerged on Academic Torrents, where it has been downloaded hundreds of times without any restrictions. MS-Celeb was also repackaged into another face dataset called "Racial Faces in the Wild" (RFW). To create it, the RFW authors uploaded face images from the MS-Celeb-1M dataset to the Face++ API and used the inferred racial scores to segregate people into four subsets: Caucasian, Asian, Indian and African, each with 3,000 subjects.

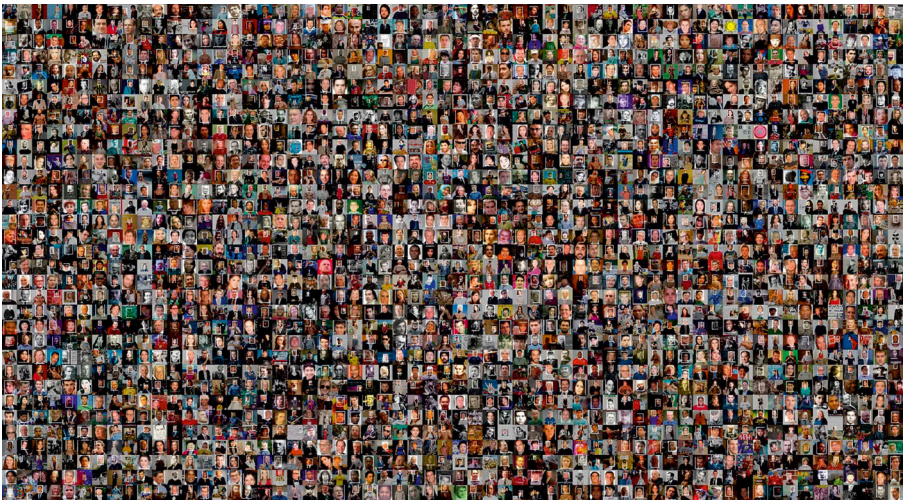

Fig. 8 A still image from MS-Celeb dataset.

Meanwhile, Microsoft researchers never actually stopped using the MS-Celeb-1M dataset. A November 2019 research paper posted to the pre-print server Arxiv entitled "A Scalable Approach for Facial Action Unit Classifier Training Using Noisy Data for Pre-Training" (Fung and McDuff 2019) cites using "the large scale publicly available MS-Celeb-1M dataset" for "the pre-training stage" of building automated facial action unit classification technology. The author of the paper is affiliated with Microsoft Research.

Conclusion

From one perspective, "in the wild" is an ideal characteristic for training data because it can provide a closer match to an unknown deployment environment. Theoretically, this can improve real-world performance by reducing disparity and bias. In reality, data collected from sources "in the wild" inherit new problems including the systemic inequalities within society and are never "natural" or "wild." Representing datasets as unconstrained or "wild" simplifies complexities in the real world where nothing is free from bias. Further, collecting data without consent forces people to unknowingly participate in experiments which may violate human rights.

However, for certain types of datasets or applications, it may be in the public interest to provide publicly accessible data. Not all datasets contain faces or biometric information. Creative Commons licenses were designed to unlock the restrictive nature of copyright and allow creators to share and remix each other's work. Allowing Creative Commons images for machine learning and artificial intelligence applications may be public utility, as the Wikimedia CEO, Ryan Merkly has noted (Merkley 2019), but only if better regulations are created to protect biometric information that can be exploited for surveillance and biometric technologies, with serious implications for privacy and human rights. Currently, Creative Commons is not interested in such a license, but our research suggests it may in the public interest to continue this pursuit, or else develop alternative data licensing schemes and move away from using Creative Commons.

Additional datasets and utilities, including a search engine to help locate social media images in datasets, are currently being developed and will be published on our research project website https://megapixels.cc.*

* Editorial note: In January 2021 MegaPixels was transferred into a new project, Exposing.ai: https://exposing.ai/.

References

Bansal, Ankan, Anirudh Nanduri, Carlos D. Castillo, Rajeev Ranjan, and R.Chellappa. 2017. "UMDFaces: An Annotated Face Dataset for Training Deep Networks." *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 464–73.

Benfold, Ben, and Ian Reid. 2011. "Stable Multi-Target Tracking in Real-Time Surveillance Video." *CVPR 2011*, 3457–64.

Berg, Tamara l., Alexander C. Berg, Jaety Edwards, Michael Maire, Ryan White, Yee-Whye Teh, Erik Learned-Miller, and D.A. Forsyth. 2004. "Names and Faces in the News." *CVPR 2004*.

Bouma, Henri, Patrick Hanckmann, Jan-Willem Marck, Leou Penning, Richard den Hollander, Johan-Martijn ten Hove, Sebastiaan P. van den Broek, Klamer Schutte, and G.J. Burghouts. 2012. "Automatic Human Action Recognition in a Scene from Visual Inputs." *Defense + Commercial Sensing*.

Burton-Harris, Victoria, and Philip Mayor. 2020. "Wrongfully Arrested Because Face Recognition Can't Tell Black People Apart." *ACLU*, June 24, 2020. https://www.aclu.org/news/privacy-technology/wrongfully-arrestedbecause-face-recognition-cant-tell-black-people-apart/.

Chavdarova, Tatjana, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. 2017. "The Wildtrack Multi-Camera Person Dataset." *ArXiv* abs/1707.09299.

Chen, Bor-Chun, Chu-Song Chen, and Winston H. Hsu. 2015. "Face Recognition and Retrieval Using Cross-Age Reference Coding with Cross-Age Celebrity Dataset." *IEEE Transactions on Multimedia* 17: 804–15.

Darby, Chris. 2019. "In-Q-Tel President Chris Darby on Investment and Innovation in U.s. Intelligence Intelligence Matters." *Intelligence Matters*, April 23, 2019, Podcast. https://podcasts.apple.com/us/podcast/in-q-tel-president-chrisdarby-on-investment-innovation/id1286906615?i=1000436184139.

Fung, Alberto, and Daniel McDuff. 2019. "A Scalable Approach for Facial Action Unit Classifier Training Using Noisy Data for Pre-Training." *ArXiv* abs/1911.05946.

Grother, Patrick J., Mei L. Ngan, and Kayee K. Hanaoka. 2019. "Face Recognition Vendor Test Part 3: Demographic Effects." *NIST Interagency/Internal Report (NISTIR),* 8280.

Guo, Yandong, and Lei Zhang. 2017. "One-Shot Face Recognition by Promoting Underrepresented Classes." *ArXiv* abs/1707.05574.

Guo, Yandong, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. "MS - Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition." *ECCV*.

Günther, Manuel, Peiyun Hu, Christian Herrmann, Chi-Ho Chan, Min Jiang, Shunchuang Yang, Akshay Raj Dhamija, et al. 2017. "Unconstrained Face Detection and Open-Set Face Recognition Challenge." *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 697–706.

Hill, Kashmir. 2019. "How Photos of Your Kids Are Powering Surveillance Technology." *New York Times*, October 10, 2019. https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html.

Hinton, Geoffrey. 2017. "Heroes of Deep Learning: Andrew Ng interviews Geoffrey Hinton." August 8, 2017. Video, 39:45. https://www.youtube.com/watch?v=-eyhCTvrEtE.

Huang, Gary B., Marwan Mattar, Tamara Berg and Eric Learned-Miller. 2008. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*. Erik Learned-Miller, Andras Ferencz and Frédéric Jurie, October 2008, Marseille, France. inria-00321923. https://hal.inria.fr/inria-00321923/document.

Human Rights Watch. 2019. "China: Police 'Big Data' Systems Violate Privacy, Target Dissent." November 19, 2017. https://www.hrw.org/news/2 017/11/19/china-police-big-data-systems-violate-privacy-target-dissent.

Islam, Mohammad T., Scott Workman, and Nathan Jacobs. 2015. "FACE2GPS: Estimating Geographic Location from Facial Features." *2015 IEEE International Conference on Image Processing (ICIP)*, 1608–12.

Jain, Vidit, Eric Learned-Miller, and Andrew McCallum. 2007. "People-LDA: Anchoring Topics to People Using Face Recognition." *2007 IEEE 11th International Conference on Computer Vision*, 1–8.

Kormann, Judith. 2020. "Wie Unsere Bilder Zu Überwachungstechnologie beitragen." *NZZ*, November 3, 2020. https://www.nzz.ch/schweiz/ueberwachungwie-unsere-bilder-die-technologie-verbessern-ld.1542751.

Lee, Justin. 2017. "PING AN Tech Facial Recognition Receives High Score in Latest LFW Test Results." February 13, 2017. https://www.biometricupdate.com/201702/ping-an-tech-facial-recognitionreceives-high-score-in-latest-lfw-test-results.

Lee, Kaifu. 2019. "Frontline: In the Age of AI." *In the Age of AI*. November 14, 2019. Podcast. https://podcasts.apple.com/de/podcast/frontline-film-audio-track-pb s/id336934080?l=en&i=1000456779283.

iBug. 2019. Website. https://ibug.doc.ic.ac.uk/resources/lightweight-face-recognition-challenge-workshop/.

Merkley, Ryan. 2019. "Use and Fair Use: Statement on Shared Images in Facial Recognition AI." *Creative Commons*, March 13, 2019. https://creativecommons.org/2019/03/ 13/statement-on-shared-images-in-facial-recognition-ai/.

Metz, Cade. 2019. "Facial Recognition Tech Is Growing Stronger, Thanks to Your Face." *New York Times*, July 13, 2019. https://www.nytimes.com/2019/07/13/technology/databases-faces-facial-recognition-technology.html.

Mozur, Paul. 2019. "One Month, 500,000 Face Scans: How China Is Using AI To Profile a Minority." *New York Times*, April 14, 2019. https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

Murgia, Madhumita. 2019. "Who's Using Your Face? The Ugly Truth About Facial Recognition." *Financial Times*, September 18, 2019. https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e.

Nech, Aaron, and Ira Kemelmacher-Shlizerman. 2017. "Level Playing Field for Million Scale Face Recognition." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3406–15.

Phillips, P. Jonathon, Hyeonjoon Moon, Patrick J. Rauss, and Saadat A. Rizvi. 1997. "The Feret Evaluation Methodology for Face-Recognition Algorithms." *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 137–43.

Ristani, Ergys, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. 2016. "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking." *ECCV Workshops*.

Rothe, Rasmus, Radu Timofte, and Luc Van Gool. 2015. "DEX: Deep Expectation of Apparent Age from a Single Image." *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 252–57.

Stewart, Russell, Mykhaylo Andriluka, and Andrew Y. Ng. 2016. "End-to-End People Detection in Crowded Scenes." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2325–33.

Xiang, Chao, Haochen Shi, Ning Li, Meng Ding, and Huiyu Zhou. 2019. "Pedestrian Detection Under Unmanned Aerial Vehicle an Improved Single-Stage Detector Based on Retinanet." *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–6.

Xu, Jing, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. 2018. "Attention-Aware Compositional Network for Person Re-Identification." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2119–28.

Yi, Dong, Zhen Lei, Shengcai Liao, and Stan Z. Li. 2014. "Learning Face Representation from Scratch." *ArXiv* abs/1411.7923.