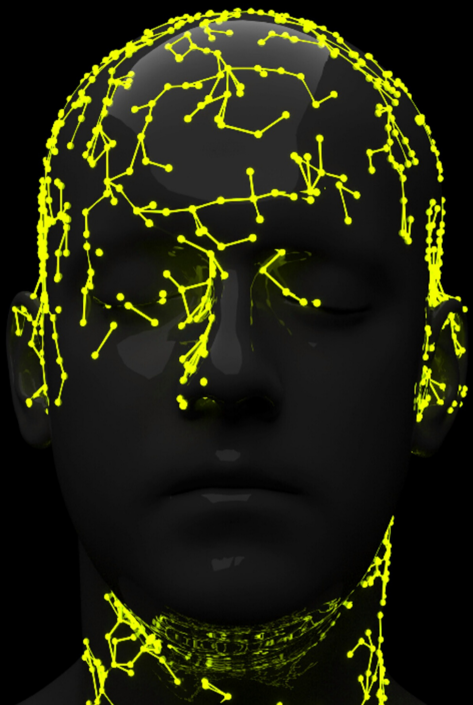
FACIAL RECOGNITION TECHNOLOGIES:

A PRIMER

Joy Buolamwini, Vicente Ordóñez, Jamie Morgenstern, and Erik Learned-Miller

May 29, 2020



Contents

Acknowledgements:
Facial Recognition Technologies in the Wild: A Call for a Federal Office and the supplemental document *Facial Recognition Technologies: A Primer* were made possible by a grant from the John D. and Catherine T. MacArthur Foundation to the Algorithmic Justice League.

We thank the following individuals for comprehensive feedback on earlier drafts of these documents: Sasha Costanza-Chock, Aaina Agarwal, Ben Hutchinson, Brant Cheikes, David Evans, and Paul Humphreys.

ALGORITHMIC JUSTICE LEAGUE MacArthur Foundation

Facial Recognition Technologies: A Primer

This primer is meant to accompany our white paper, *Facial Recognition Technologies in the Wild: A Call for a Federal Office*, as a supporting document. It presents background on Facial Recognition Technologies (FRTs) and provides important context for material in the main document of the white paper.

The primer is written for a non-technical audience to increase understanding of the terminology, applications, and difficulties of evaluating this complex set of technologies. In Section 1, we provide basic definitions of common terms like *face detection* and *face verification.* Such definitions are needed to clarify the precise meaning of subsequent discussions. In Section 2, we present some common and lesser known uses of FRTs. Section 3 introduces some of the fundamental technical concepts used in the process of recognizing a face, from the capture of the face by a camera, to the digital representation of faces in a computer, and finally to the evaluation of results and the categorization of errors. Section 4 highlights challenges with characterizing and measuring the accuracy of FRTs. This primer is a basic tutorial and does not provide guidance on if, how, or when specific FRTs should be used.

1. What are Facial Recognition Technologies (FRTs)?

We define *Facial Recognition Technologies* (FRTs) to be a set of digital tools used to perform tasks on images or videos of human faces. These tools can be grouped into three broad categories depending upon the question they answer:

1. **Is there a face in the image?**
2. **What kind of face is shown in the image?**
3. **Whose face is shown in the image?**

These questions cover many of the common use cases of FRTs. For each of the questions above, we discuss applications that answer the question in a particular way.

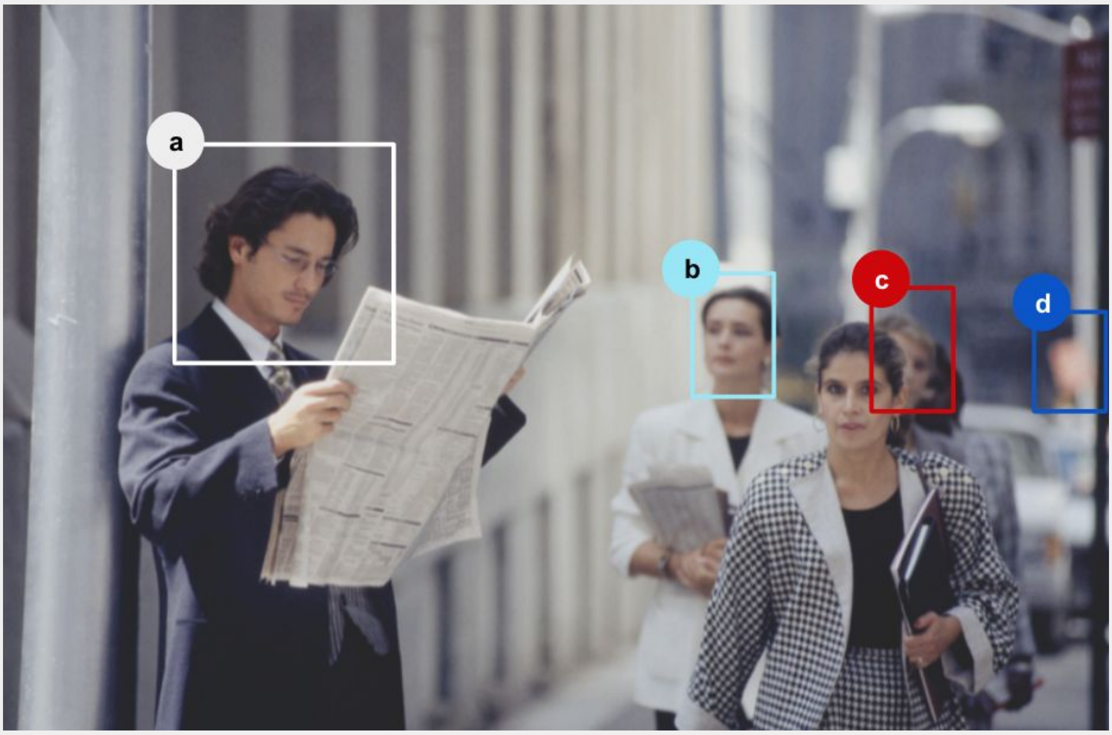
1.1 Is there a face in the image?

- **Face detection.**[1] Face detection is the process of detecting the presence of faces and locating those faces in an image or video (see Figure 1). Detecting the presence of a face and locating it in the image is not the same as assigning a unique identity to a

[1] The term *facial detection* is emerging in use but the vast majority of published research uses the term face detection.

detected face or trying to determine attributes like gender or age. In particular, the process of face detection does not report anything about who someone is or what kind of person someone might be. It is merely the process of attempting to find and locate faces in an image. Subsequent analysis performed on a face often depends on the successful completion of face detection.

Figure 1: Challenges of face detection.



This figure shows an image in which some, but not all of the faces have been detected and indicated with boxes. Deciding whether or not a region represents a face can be made challenging due to viewing angle (box a), image blur (box b), partial occlusion (box c), and many other factors. False positives may also be detected (box d). Some face detectors have been shown to perform poorly on people with darker skin tones [12].

Figure 1 shows some examples of detected faces in an image, and some of the phenomena that make face detection challenging. Note that a face detection system may report zero faces, one face, or many faces in an image.

Face detection software can make two kinds of errors. It might fail to find a face that is present, such as the face of the person in the checked jacket in Figure 1. This is called a *false negative.* The other type of error is when a face detector identifies a non-face structure as a real face. For example, in Figure 1, box d shows the incorrect detection of an object in the background that is not a face. This is called a *false positive*. Now we turn to the process of characterizing properties or attributes of faces.

1.2 What kind of face is shown in the image?

Another major category of FRTs are applications that try to assess something about a person other than their identity, such as their gender, age, or their emotional state. We break this down into two sub-categories.

- **Face attribute classification** and **face attribute estimation.** Software can be developed to assess the attributes of a person from their face (see Figure 2). When these attributes have been separated into distinct categories,[2] such as gender, race, or ethnicity, this may be called face attribute classification. When the attribute is a number, like an age, the term face attribute estimation is more appropriate. The terms *gender classification* and *age estimation* also appear frequently in the scientific literature and popular press, referring to specific instances of these techniques. Finally, software to detect and locate accessories like glasses and scarves or face attributes like beards or mustaches may be referred to as **face attribute detection.**

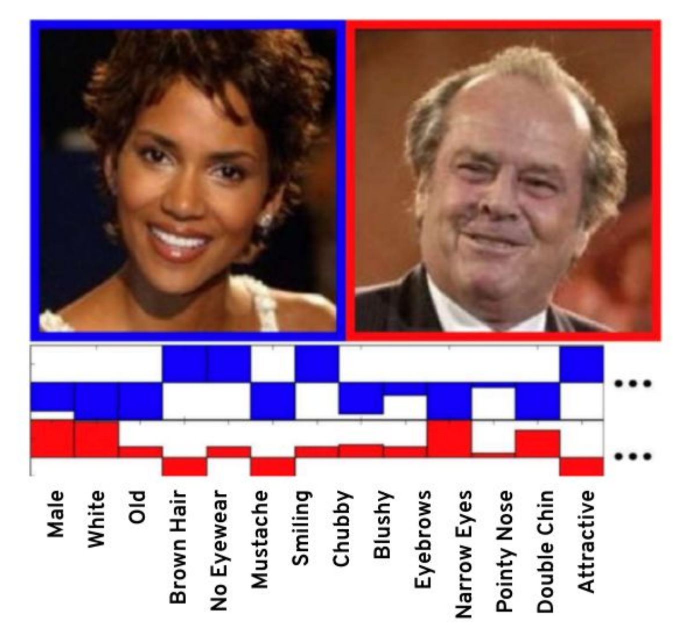


Figure 2: Estimation of facial attributes.

This figure, reproduced from Kumar et al. [24], illustrates how FRTs may estimate facial attributes defined by their developers. The estimation of each attribute is indicated with a rectangle: above the central line indicates the likely presence of the feature, and below the line the absence of the feature. The goal of the underlying FRT is to mimic subjective human assessments about these attributes.

- **Emotion**, **affect**, and **facial expression classification**. Facial recognition technologies can be used to classify facial expressions, such as “smile,” “frown,” or “scowl.” They can also be used for the closely related problem of inferring the emotional state or *affect* of a person, such as “happy,” “sad,” or “angry.”

[2] The authors acknowledge that many terms such as gender, race, and ethnicity, are socially constructed categories, differ across societies, cultures, and over time, and have no universally accepted meaning. Nevertheless, practitioners may attempt to categorize individuals into groups such as binary ’male’ and ’female’ based on their own notions of categories.

A good deal of confusion arises from the ambiguous use of these terms in the scientific literature. In particular, many systems that claim to be doing "emotion recognition" would be better described as doing "facial expression recognition." That is, when a facial recognition system reports "happy" as a label for a face, in most cases this refers to an expression like a smile, not to the true emotional state of the individual. It is important to keep in mind that many systems that claim to do emotion recognition have really been developed to recognize specific facial expressions (as performed by paid actors), not to detect the subtle cues that may reveal a person's underlying emotional state.[3]

In the accompanying white paper, we delve further into the need for evidence to support claims that are made concerning the attributes FRTs can derive from a face.

1.3 Whose face is shown in the image?

The final category of applications is related to establishing the identity of a person, or to the related question of whether two pictures represent the same person.

- **Face recognition** or **facial recognition**. These terms are often used informally as a catchall phrase, referring to a wide variety of processes. However, the more precise and specific meaning used by researchers and developers is the process of using digital representations of faces to try to identify or verify the identity of a unique individual. The image of a particular individual we wish to recognize is often referred to as the *query image* or a *query.*[4]

  There are two subtly different types of recognition, referred to as **face verification** and **face identification**, which are defined below. Note that *recognition* differs from *detection*, the latter of which focuses solely on the presence of a face, and not its identity.

- **Face verification** or **facial verification**. Face verification is one type of face recognition. It attempts to determine whether an image shows a particular person. For example, software on a cell phone may try to answer the question, "Can it be **verified** that the camera shows the phone's owner?" A query image is deemed to be either a *match*, if it appears to show the owner, or a *mismatch* otherwise.

  There are two common ways to perform face verification. In the first, one asks a question such as "Does this image show Janelle Smith?", in which the person of interest is named. To answer this question, a system needs some prior source of information about the appearance of Janelle Smith, such as previously obtained pictures or a description. A common use for this type of face verification is *access control*, such as software that allows the owner of a device or a service to access it. Access control can be used to unlock a phone, access a bank account, or pay for an item with a digital currency. If you use face verification to access your cell phone, the face verification system takes a new picture of your face (the query image) and compares it to information it obtained previously to try to assess whether you are the same person (see more details about this process in Section 3).

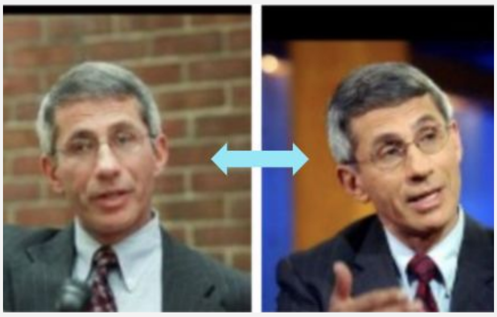
[3] For a comprehensive survey on limitations of inferring emotions from facial expressions see the article by Barrett et al. [10] or the on-line discussion of it [2].

[4] The terms *probe image* and *probe* are also widely used in industry and research.

In the second common version of face verification, one is given two pictures and asks, "Is the first person the same as the second person?" An example of such face verification is shown in Figure 3 (the two pictures show the same person). In this case, it is not necessary to know the identity of either person to answer the question. Face verification is also referred to as *1-to-1 matching* or *1-to-1 comparison*.

Like other facial recognition technologies, face verification systems will inevitably make errors. The two types of errors are typically referred to as a *false match*, in which the system incorrectly reports that face images of two different people are the same, and a *false mismatch*, in which the system incorrectly reports that face images of the same person are different.

Figure 3: Face verification.



The task of face verification asks, "Is the person on the right the same as the person on the left?" This task may be challenging due to the similarity in appearance of two people, or due to imaging factors like lighting, blur, or image quality. This pair of images, which shows the same person, was taken from *Labeled Faces in the Wild*, a popular benchmark for face verification [23].

- **Face identification** or **facial identification**. Face identification is the second major type of face recognition. Face identification attempts to answer the question, "Whose face is this?" Face identification software can only match the image of a face to a person for whom it already has some appearance information. The set of people for whom an application has stored appearance information is called the *gallery*. Simply put, this is the set of people that a face identification system could possibly identify. A typical example of a gallery would be the set of people who work in a secured location, such as a private office building. The correct answer to a face identification query is either the identity of a person in the gallery (e.g., "Employee #347") or "none of the above" if the image shows a person who is not in the gallery. Face identification can be used for surveillance, to find a person of interest, or for the identification of subjects that are either unable or unwilling to respond. Face identification goes by many different names. It may be referred to as *1-to-many comparison*, *1-to-many matching, 1-to-many identification,* or *1-to-N identification.*

We believe that understanding of these widely used terms will help the reader of our white paper better understand the risks, benefits, and implications of facial recognition technologies. In the next section, we introduce both some common and some lesser known applications of FRTs.

2. How and Where are FRTs Used?

Facial recognition technologies are quickly entering many aspects of everyday life around the world. This section presents a non-exhaustive overview of different domains in which the technologies are being applied.

Banks
Financial institutions are using face verification to add security for banking transactions [3].

Consumer Products
Devices like laptops and phones can be accessed using face verification. Digital applications allow users to apply digital filters mapped to facial features. Social media websites can automatically identify and tag individuals in images and videos.

Events
Entertainment events like sporting events and concerts can use face verification for ticketing. Public gatherings for entertainment or other purposes like community events or protests can be subject to real-time face identification [17, 16, 4].

Housing
Individual homeowners can install camera systems that integrate FRT in consumer products like smart doorbells. Landlords can install face recognition entry systems for tenants to enter buildings [5, 18].

Police Departments
Law enforcement officers can use face recognition to support investigations, search mugshot databases, or attempt to identify individuals who are uncooperative or incapacitated.

Places of Worship
Some churches are using face recognition to track congregation attendance [11,7,22].

Schools
From taking attendance to attempting to assess student attentiveness, FRTs are being introduced in school districts around the world [8, 9, 15].

Stores
Retailers are integrating face-based payments to allow customers to pay for items with just a look into a camera [6, 26]. With cameras embedded in shelves, mannequins, or more traditionally on ceilings, stores can use FRT to assess the demographics of shoppers for marketing purposes or block entry into a store if an individual is flagged as suspect [20, 27].

Transportation
As ports of entry, airports have become a testing ground for face verification, where it is being used to verify passport IDs. Face-based access can be used by patrons to access public transportation. In New York City, detected faces are displayed on monitors, reportedly to deter fare dodgers [25]. Car manufacturers are integrating these technologies to allow drivers to access their cars and also monitor drivers for warning signs of drowsiness or inattentiveness [1].

Workplaces
Employers can use face identification to limit access of work spaces to employees [13]. Others are using facial analysis on videos of job candidates to inform hiring decisions [9, 14].

It is important to remember that in any of these domains, there is a wide range of possible applications of facial recognition technologies, from the detection of a face without recognition (say, for counting customers), to the identification of specific people, or even to the long term storage and surveillance of citizens. Each of these capabilities may be performed either with or without the consent of the individual.

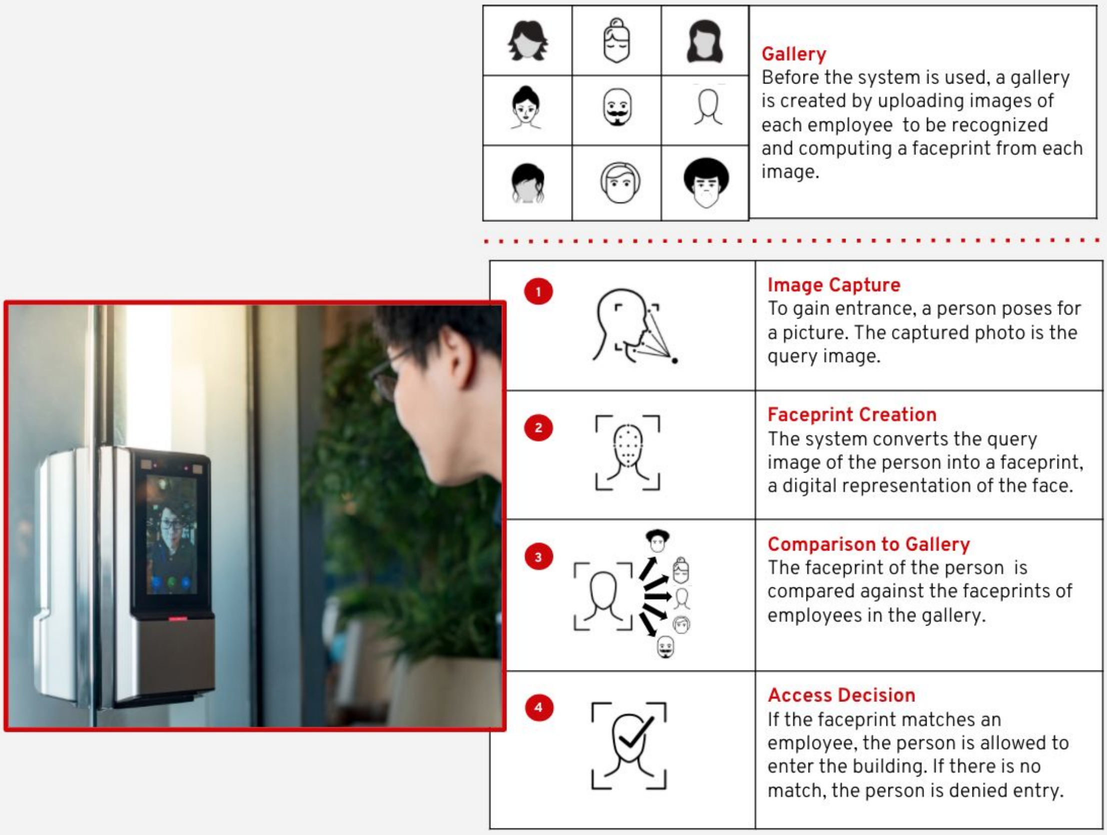
With a sense of some of the standard terminology and applications of FRTs, we now move to a description of key processes used in recognizing an individual face.

3. How Does a Machine Recognize an Individual Face?

Face recognition is a *biometric* technology. A biometric technology uses automated processes to recognize an individual through unique physical characteristics (fingerprints, iris pattern) or behaviors (walking motion, speech patterns). Face recognition is notable as a biometric technology since it can be used from a distance (unlike fingerprints, for example), and because it can be performed without the knowledge of the individual. Another critical aspect of biometrics is their reliability. Because the accuracy of face recognition is heavily dependent upon the conditions of use, its use as a biometric is further complicated.

To understand the recommendations in our main report, it is important to have some understanding of how facial recognition technologies work. Many face recognition systems share similar components. Below we discuss five of the common components found in many (but not all) face recognition systems.

Figure 4: Face identification for workplace access.



Gallery
Before the system is used, a gallery is created by uploading images of each employee to be recognized and computing a faceprint from each image.

1 **Image Capture**
To gain entrance, a person poses for a picture. The captured photo is the query image.

2 **Faceprint Creation**
The system converts the query image of the person into a faceprint, a digital representation of the face.

3 **Comparison to Gallery**
The faceprint of the person is compared against the faceprints of employees in the gallery.

4 **Access Decision**
If the faceprint matches an employee, the person is allowed to enter the building. If there is no match, the person is denied entry.

The figure shows the process of using a face identification system for limiting access to a building to a set of employees.

3.1 Capture and detection

To be used in a face recognition system, a face must be photographed by a camera or recorded by a video camera. One primary use of such photos is to build a *gallery* of people to be recognized (see Section 1.3). The second major use is at recognition time, when one is trying to identify or verify the identity of a person, by comparing an image of that person to the gallery. In capturing photos for face recognition, a particular application may require that the subject poses in a specified position. Such requirements are typically used for acquiring photos for driver's licenses or passports. In this case, the face will appear at a known location in the middle of the photo. On the other hand, if a photo is captured in an unstructured setting, such as in surveillance footage, the face may need to be located with face *detection* software in order to be further analyzed (see Section 1).

The process of having one's face photographed can be voluntary, as when a person attempts to unlock a phone with their face. But it can also be involuntary, such as when an image posted online is selected for recognition purposes. Any image or video containing a face can serve as a starting point for recognition with or without notice and consent.

3.2 Enrollment

For a face recognition system to verify or identify an individual, prior facial information needs to be provided. In order to build a system for face identification, one must gather information about the appearance of a set of individuals, so that later, a new image can be compared to this information. The set of people for whom visual information has been gathered is referred to as the *gallery* (see Section 1). The process of recording visual information about an individual for inclusion in the gallery is called the *enrollment* of that individual. For example, if a new employee is hired at a company that uses face-based access control, the employee may have to be enrolled in a gallery by having a photo or series of photos taken of their face and added to a gallery database. For face verification used by an individual on a consumer device like a phone, enrollment occurs when the user configures the system by uploading an initial image of their face.

3.3 The digital representation of a face

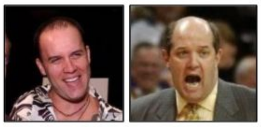
Once an image is taken and a face is detected in the image, characteristics of the face may be stored in a numerical format called a *template* or *faceprint*. A digital template or faceprint is analogous to the digital representation of a fingerprint. Software developers strive to make faceprints that are as informative as possible about identity. In the development of facial recognition technology by a developer or researcher, the goal is for faceprints generated from two pictures of the same person to be highly similar, while faceprints from two different individuals should be significantly different. This goal is not easy to achieve, however. A simple example of this difficulty is illustrated by the challenge of making faceprints different for identical twins.

3.4 Comparison

Once a faceprint has been extracted from the image of a face, the comparison process can start.

- **Face verification: 1-to-1 comparison.** If the goal is to confirm the identity of a face, for example to access a bank account, the faceprint is compared to an existing faceprint of the expected individual. The individual makes a claim of who they are, and the task of the system is to verify this claim.
- **Face identification: 1-to-many comparison.** If the goal is to identify a face, for example in surveillance footage, the faceprint of interest is compared not just to one faceprint but to the whole set of faceprints in the gallery. Unlike verification, the task of identification requires searching a gallery of potential matches.
- **Similarity scores** or **match scores.** When two faceprints are compared by a machine, a similarity score may be computed to represent the similarity of the two faceprints (see Figure 5).

Figure 5: Error trade-offs in face recognition.

	IMAGE PAIR	SIMILARITY SCORE	SIMILARITY SCORE THRESHOLD FOR MATCH			
			60	70	80	90
MISMATCH		65	✗ Match	✓ Mismatch	✓ Mismatch	✓ Mismatch
MATCH		73	✓ Match	✓ Match	✗ Mismatch	✗ Mismatch
MATCH		83	✓ Match	✓ Match	✓ Match	✗ Mismatch
MISMATCH		85	✗ Match	✗ Match	✗ Match	✓ Mismatch
MATCH		95	✓ Match	✓ Match	✓ Match	✓ Match
	Total False Matches (False Positives)		2	1	1	0
	Total False Mismatches (False Negatives)		0	0	1	2
	Total Error Rate		2/5	1/5	2/5	2/5

This figure illustrates the trade-off that must frequently be made between two types of errors in face verification. The goal of verification (see Section 3) is to decide whether two images show the same person (a *match)* or show two different people (a m*ismatch*). The correct answer is shown to the left of each image pair. The table shows how a face verification system makes different decisions (shown as "Match" or "Mismatch") by comparing the similarity score to a threshold. The green check marks indicate that the system's response is correct. The red x's show that the system has made an error. Note that as the threshold goes from 60 to 90, the number of false matches goes down, but the number of false mismatches goes up. In this example, and in most real-world scenarios, there is no single setting of the threshold which eliminates all errors. Typically, one must make a compromise by setting a threshold which balances the number of false matches and false mismatches.

Different techniques can be used to arrive at this score, so the score for one face recognition system cannot always be reliably compared to another. Since a particular person's appearance may vary significantly from one time to another, two faceprints of the same person are rarely exactly the same. Variations in many factors, such as hairstyle, camera angle, image resolution, lighting, and make-up, can all have significant impacts on faceprints, resulting in the faceprints of a single individual having significant variability. Conversely, two different people with similar superficial features (say, a certain style of beard), or whose photos were taken under similar conditions may, in some cases, have nearly identical faceprints. Thus, highly similar faceprints do not always indicate the same individual, and significantly different faceprints do not always indicate different individuals.

Figure 5 shows five pairs of images and a similarity score for each. The figure demonstrates that the similarity score for two different people may sometimes be higher than the similarity score for two pictures of the same person. **The inability for any technology to generate a unique faceprint for each individual is at the heart of many face recognition system errors.** We can say, however, that generally speaking, the higher the similarity score the more likely the faceprints being compared belong to the same individual.

3.5 Matching decision

In *face verification* (1-to-1 comparison), the goal is to decide whether two images match (i.e., whether they represent the same person). Figure 5 shows five pairs of images that are part of a popular face verification benchmark.[5] The label to the left of each image pair indicates whether the images show the same person (match) or not (mismatch).

Comparing the faceprints for each person in the pair results in a similarity score, as described above. Because there is uncertainty about this similarity, a threshold or cut-off point can be set in order to decide whether the two faceprints match or not. If a system returns a similarity score between 0 and 100 and a threshold of 80 is set, only faceprints with similarity scores at or above 80 are considered a match. The system's answers may be correct or incorrect, and will vary as a function of the similarity threshold.

In Figure 5, the second column shows a hypothetical similarity score for each pair of images. Columns 3-6 show the decision of match or mismatch made for each pair of images using the threshold at the top of the column (60, 70, 80 or 90). If the similarity score is higher than the threshold, the face verification system guesses *match*. Otherwise it guesses *mismatch*. The text in each column of the table indicates the guess made by the face verification system. A green check mark indicates that the system got the correct answer, a red 'x' that the system made an error. **Notice that there is no setting of the threshold for which the face verification system gets all of the answers correct (no column is error free).**

In *face identification* (1-to-many comparison), a query face is compared not just to one other face, but to the entire gallery which has been obtained (see Section 1).

[5] The purpose of this face verification benchmark is to provide a standardized test for developers of face verification systems. Such benchmarks are discussed at length in our main document.

In comparing a query face to an entire gallery, one may obtain a variety of results:

- zero matches, indicating the given face was not a match to anyone in the gallery;
- one match, indicating the given face matches exactly one person in the gallery;
- multiple matches, indicating that the query face matched multiple people in the gallery.

In the case of multiple matches, one needs an additional procedure to evaluate the multiple matches and decide which of the multiple matches, if any, is correct. Often (but not always), this second phase of the problem is referred to a human being (rather than an automated facial recognition system) for a final decision.

Next, we discuss the terminology around correct and incorrect answers from a face recognition system.

- **True positive** (or **true match**). In face verification (1-to-1 comparison), a true positive (or true match) occurs if a query image matches a specific identity in a 1-to-1 comparison. For face identification (1-to-many comparison), a system may return matches to multiple people in the gallery. Clearly, at most one of these can be a true positive. As mentioned above, when multiple matches are made, a human decision maker may be consulted to compare the face of interest with the matches returned.

- **True negative** (or **true mismatch**). In addition to verifying and identifying a unique individual, systems should also correctly reject faces that do not match. For facial verification systems, a true negative means that an imposter is unable to pass themselves as somebody else. That is, if a query image (showing Alice) is compared to an enrolled person Bob, then the system's determination that they are different would be a true negative.

  For face identification systems that can return a set of possible matches, there is a possibility that the query face is not contained in the gallery. That is, the system is being presented with a picture of someone for whom it has no visual information. In this case, success means that the system does not return any matches to the query. For example, if John Doe is checked against a gallery of possible identities in which he is not present, a true negative would indicate that the system returned zero matches.

- **False positive** (or **false match**). A false positive means the wrong person is deemed to be a match. Depending on the application, the consequences of such an incorrect decision can vary. For access to a bank account using facial verification, a false positive means an imposter can access sensitive information and make unapproved transactions. For facial identification, a false positive or misidentification can occur when a system returns a list of potential candidates and the wrong person is selected by a human operator. This can occur either because the true match was not in the options returned or because the human operator missed the true candidate in the list of returned options. For processes that do not use a human operator, a misidentification can occur when a system automatically assigns the wrong identity to a person. For example, if a store is using surveillance to monitor for known shoplifters, a shopper could be incorrectly matched to a shoplifter on a hotlist and subjected to unwarranted scrutiny or false arrest.

- **False negative** (or **false mismatch**). Rejecting the correct person results in a false negative outcome (or false mismatch). For facial verification used for fraud detection, a false negative can mean an individual is denied access to a service or opportunity. For example, if ride-share companies use facial verification to establish driver identity, false negatives could prevent legitimate drivers from participating in economic opportunity. For facial identification used by law enforcement, a false negative can result in a person of interest being overlooked. For example, a missing person who may appear on footage obtained by police can be overlooked. Similarly, a suspect in an investigation could also be missed.

The choice of a matching threshold or cut-off point comes with trade-offs that must be considered depending on the application. The trade-offs include balancing success and failure modes. For example with facial identification, setting a high threshold or cut-off point reduces the likelihood of misidentifications; however a high threshold also increases the likelihood of false negatives where a person of interest is missed. Because there are trade-offs associated with different threshold settings, it is not very meaningful to characterize the accuracy of a facial identification system with a single number.

To complicate matters further, a system will often perform differently in different settings and on different sub-populations. Factors such as lighting, image quality, and camera motion can have dramatic effects on the results of a face recognition system. The December 2019 report from the National Institute of Standards and Technology (NIST) on demographic effects in facial identification uncovered that false positive rates varied across demographics by factors of up to 100. In many experiments, the highest false positive rates[6] were in West African, East African and East Asian people. The lowest false positive rates[7] were on Eastern European individuals [21]. The issues arising with choosing a threshold and different demographic performance can compound one another, as one threshold setting may lead to better results for one group and poorer outcomes for another. In the last section, we discuss further the issues and challenges around assessing the accuracy of FRTs.

4. How Accurate are FRTs?

A commonly asked question is *how accurate is facial recognition*?

Though the question seems simple, there are many ways in which accuracy can be defined. These different definitions may lead to varying conclusions about the same system. In addition, when facial recognition is used as a catchall phrase by the media or in policy discussions, we must keep in mind that there are many different kinds of facial recognition technologies - from systems that detect the mere presence of a face, to those that assign attributes to a face, and finally those that attempt to verify or identify a unique individual (see Section 1).

[6] Differences in false positive rates are typically measured by setting a similarity score threshold for a fixed false negative rate, and then examining the resulting false positives.

[7] Similarly, false positive rates are measured by setting a threshold to achieve a given false negative rate, and then examining the resulting false positives.

In addition, for a specific set of FRTs like those focused on face verification, systems produced by different companies will produce different accuracy results. Though it is tempting to try to understand how well a specific system works by focusing on a single accuracy number, we must consider the different types of errors a system makes, the distribution of those errors across different demographic populations, and how real-world conditions differ from test conditions.

4.1 Performance metrics and benchmarks

We focus on performance to emphasize there isn't just one way to assess how well FRTs work. Like with cars, a number of attributes factor into assessing the performance of FRTs. Depending on how a specific system is used, those attributes hold different weight. Statistics such as the number of false matches and false mismatches are often used to characterize the performance of a face recognition system. For example, for the five pairs of images in Figure 5, a face verification system that produces the given similarity scores and uses a threshold of 90 (shown in the rightmost column) would produce zero false matches and two false mismatches. These statistics, when measured over large, standardized collections of images are referred to as *performance metric*s, and the datasets for computing such metrics are called *benchmarks.* Our accompanying white paper contains an appendix dedicated exclusively to analyzing benchmarks and why they alone cannot provide a complete solution for determining the expected performance of a specific facial recognition technology in a real deployment.

4.2 Real-world performance and benchmark results

The most common pitfall in measuring the performance of FRTs is to assume that the statistics measured on a standard benchmark will be representative of the system's performance in real-world scenarios. For example, FRTs in recent benchmark tests, when tested on images of faces belonging to a diverse population of individuals from several countries and age groups, had very low false match rates (about 1 false match in 10,000 trials). These same FRTs, when tested on a more homogeneous population, where the matching is performed across people in the same age group and geographical area, had false match rates about 20 times higher [21]. Why does this large discrepancy occur? Because when people are the same age and share more traits, they can be harder to distinguish. The performance of a system varies with the diversity of the population on which it is applied. This illustrates again the problem with characterizing the accuracy of a face recognition system with a single number – the performance of a system is highly dependent upon the circumstances in which it is deployed. If such a face recognition system is deployed at a company where the population is similarly homogeneous, the false match rate obtained in the benchmark under a global population is almost irrelevant, and should not be used in isolation to justify the use of the system.

In addition, **seemingly small error rates can still have a negative impact on a substantial number of individuals**. For example, suppose that a system with a false match rate of 1 in 500 is deployed. If such a system is used to control access to resources (e.g., buildings) in companies across a city with a working population of 2 million people, this would result in approximately 4, 000 false matches per day. Moreover, the false matches are likely to occur unevenly for various segments of the population.

In addition to differences in the populations used for testing and in real-world deployments, differences in other conditions, such as image quality, facial orientation, occlusion, lighting, and other factors can cause the differences in benchmark results and real-world deployments to be dramatic. The question remains, what are effective alternatives to using benchmarks and metrics in order to decide if a specific facial recognition technology is appropriate for deployment for a particular application in a targeted population? Further, we posit that questions beyond accuracy and technical considerations need to be incorporated into this process that deal with issues such as harmful discrimination, privacy, consent and legality. In some cases, in certain contexts or for particular applications, the use of FRTs will not be justified regardless of accuracy. In other cases, their use should be contingent upon careful regulation and oversight. Our accompanying white paper develops our proposal of a framework and entity to handle these types of decisions in the realm of FRTs.

References

[1] Affectiva automotive AI for driver monitoring systems. https://affectiva.com/product/affectiva-automotive-ai-for-driver-monitoring-solutions/, last accessed on May 1, 2020.

[2] Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. https://psychologicalscience.org/publications/emotional-expressions-reconsidered-challenges-to-inferring-emotion-from-human-facial-movements.html, last accessed on May 1, 2020.

[3] Face recognition for financial institutions. https://findface.pro/en/solution/finance/, last accessed on May 1, 2020.

[4] Five ways facial recognition is being used in sports today. https://facefirst.com/blog/ways-facial-recognition-is-being-used-in-sports-today/, last accessed on May 1, 2020.

[5] Know who is at your doorstep with instant face recognition. https:// wisenetlife.com/en-us/product/SmartCam/SNH-V6435DN/feature/, last accessed on May 1, 2020.

[6] Smile-to-pay: Chinese shoppers turn to facial payment technology. *The Guardian*, September 4, 2019. https://theguardian.com/world/2019/sep/04/smile-to-pay-chinese-shoppers-turn-to-facial-payment-technology.

[7] Brazilian churches start to introduce facial recognition in their worship services. *Evangelical Focus*, February 5, 2020. https://evangelicalfocus.com/science/5088/Brazilian_churches_start_to_introduce_facial_recognition_in_their_services_.

[8] Davey Alba. Facial recognition moves into a new front: Schools. *New York Times*, February 6, 2020. https://nytimes.com/2020/02/06/business/facial-Recognition-schools. html.

[9] Sophia Ankel and Alba Asenjo. A school in Sweden has been fined over $20,000 for using facial recognition software to control student attendance. *Business Insider*, August 29, 2019. https://businessinsider.com/a-school-used-facial-recognition-to-illegally-record-class-attendance-2019-8.

[10] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[11] Lane Brown. There will be no turning back on facial recognition. *New York Intelligencer*, November 12, 2019. https://nymag.com/intelligencer/2019/11/the-future-of-facial-recognition-in-america.html.

[12] Joy Buolamwini. How I'm fighting bias in algorithms (Video file). https://ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms, November 2019.

[13] Chris Burt. Intel implements facial recognition security system at offices to identify threats. *Biometric Update*, May 12, 2020. https://biometricupdate.com/202003/intel-implements-facial-recognition-security-system-at-offices-to-identify-threats.

[14] Angela Chen. The AI hiring industry is under scrutiny—but it'll be hard to fix. *MIT Technology Review*, November 7, 2019. https://technologyreview.com/2019/11/07/75194/hirevue-ai-automated-hiring-discrimination-ftc-epic-bias/.

[15] Neil Connor. Chinese school uses facial recognition to monitor student attention in class. *The Telegraph*, May 17, 2018. https://telegraph.co.uk/news/2018/05/17/chinese-school-uses-facial-recognition-monitor-student-attention/.

[16] Sopan Deb and Natasha Singer. Taylor Swift said to use facial recognition to identify stalkers. *The New York Times*, December 13, 2018. https://nytimes.com/2018/12/13/arts/ music/taylor-swift-facial-recognition.html.

[17] Kevin Draper. Madison Square Garden has used face-scanning technology on customers. *The New York Times*, March 3, 2018. https://nytimes.com/2018/03/13/sports/facial-recognition-madison-square-garden.html.

[18] Lola Fadulu. Facial recognition technology in public housing prompts backlash. *The New York Times*, September 24, 2019. https://nytimes.com/2019/09/24/us/politics/ facial-recognition-technology-housing.html.

[19] Richard Feloni. I tried the software that uses AI to scan job applicants for companies like Goldman Sachs and Unilever before meeting them — and it's not as creepy as it sounds. *Business Insider*, August 23, 2017. https://businessinsider.com/hirevue-ai-powered-job-interview-platform-2017-8.

[20] Sidney Fussell. Now your groceries see you, too. *The Atlantic*, January 25, 2019. https://theatlantic.com/technology/archive/2019/01/walgreens-tests-new-smart-coolers/581248/.

[21] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (FRVT) part 3: Demographic effects. *National Institute of Standards and Technology*, 2019.

[22] Pierre Hamdi. Videos show how China has installed facial recognition scanners in Uighur mosques. *The Observers*, September 13, 2019. https://observers.france24.com/en/20190913-videos-show-how-china-has-installed-facial-recognition-scanners-uighur-mosques.

[23] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[24] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Describable visual attributes for face verification and image search. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 33, pages 1962–1977, October 2011.

[25] Adi Robertson. NYC subway denies using 'real-time face recognition screens' in Times Square. *The Verge*, April 19, 2019. https://theverge.com/2019/4/19/18507552/mta-nyc-subway-times-square-fare-evasion-cameras-facial-recognition.

[26] Alicja Siekierska. Pay with your face system coming to Canada, but not everyone is on board. *Yahoo! Finance*, October 29, 2019. https://ca.finance.yahoo.com/news/pay-with-your-face-coming-to-canada-not-everyone-on-board-142613931.html.

[27] Michael Spears. 'Look at camera for entry': Tacoma convenience store using facial recognition technology. *Kiro 7 News*, May 21, 2019. https://kiro7.com/news/south-sound-news/tacoma-convenience-store-uses-facial-recognition-technology/950979811/.