Case Study

Criminal machine learning

Criminal machine learning

For those who prefer video, this case study is described in the April 26th (https://www.youtube.com/watch?v=rga2-d1oi30) lecture of our Spring 2017 course.

In November of 2016, engineering researchers Xiaolin Wu and Xi Zhang posted an article entitled "Automated Inference on Criminality using Face Images" to a widely used online repository of research papers known as the arXiv. In their article, Wu and Zhang explore the use of machine learning to detect features of the human face that are associated with "criminality"—and they claim to have developed algorithms that can use a simple headshot to distinguish criminals from non-criminals with high accuracy. If this strikes you as frighteningly close to Philip K. Dick's notion of pre-crime, the film Minority Report, and other dystopian science fiction, you're not alone. The media thought so, too. A number of technology-focused press outlets [1 (https://www.technologyreview.com/s/602955/neural-network-learns-to-identify-criminals-by-their-faces/),2 (https://motherboard.vice.com/en_us/article/new-program-decides-criminality-from-facial-features),3 (https://theintercept.com/2016/11/18/troubling-study-says-artificial-intelligence-can-predict-who-will-be-criminals-based-on-facial-features/)] picked up on the story and explored the ethical implications.

If one could really detect criminality from the structure of a person's face, we would have an enormous ethical challenge. How would we have to adjust our notions of inalienable individual rights once we had the ability identify people as criminals before they ever acted?

Cesare Lombroso's physiognomic criminology

In the 19th century, an Italian doctor named Cesare Lombroso studied the anatomy of hundreds of criminals in an effort to develop a scientific theory of criminality. He proposed that criminals were born as such, and that they exhibit both psychological drives and physical features that harken back to what were, in his view, the subhuman beasts of our deep evolutionary past. Lombroso was particularly interested in what could be learned from facial features. In his view, the shape of the jaw, the slope of the forehead, the size of the eyes, and the

structure of the ear all contained important clues about a man's moral composition. None of this turned out to have a sound scientific basis. Lombroso's theories — many of which wrapped racist ideas of the time in a thin veneer of scientific language — were debunked in the first half of the 20th century and disappeared from the field of criminology which Lombroso helped to found.

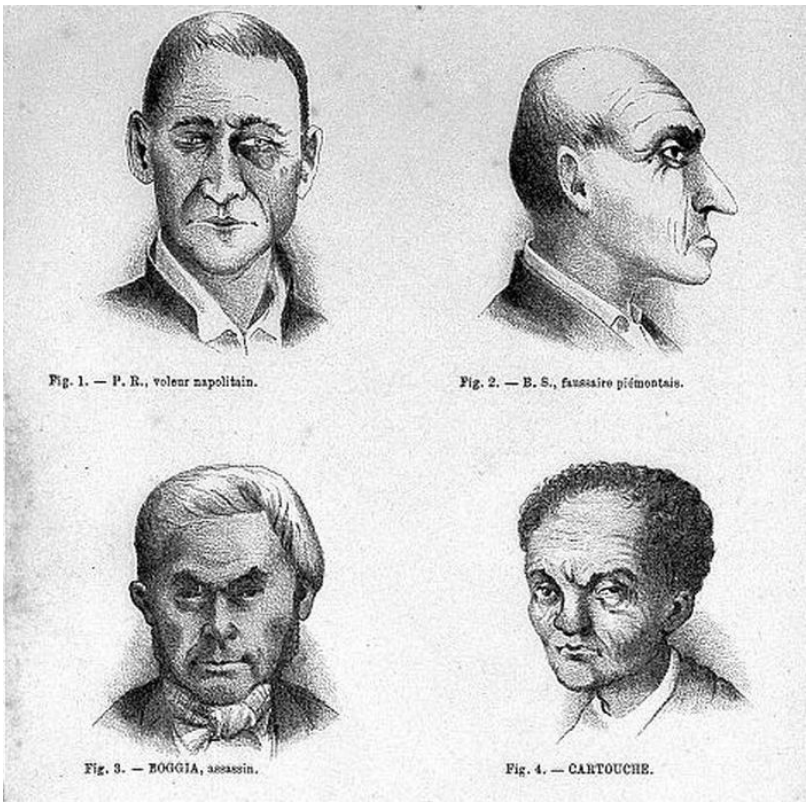

Figure 1. Criminal faces from Cesare Lombroso's 1876 book *L'Homme Criminel*

Wu and Zhang's (2016) approach

In their 2016 paper, Wu and Zhang revisit Lombroso's program. Essentially, they aim to determine whether advanced machine learning approaches to image processing can reveal subtle cues and patterns that Lombroso and his followers could easily have missed. To test this hypothesis, the authors deploy a variety of machine learning algorithms in a misguided physiognomic effort to determine what features of the human face are associated with "criminality".

Wu and Zhang claim that based on a simple headshot, their programs can distinguish criminal from non-criminal faces with nearly 90% accuracy. Moreover, they argue that their computer algorithms are free from the myriad biases and prejudices that cloud human judgment:

> "*Unlike a human examiner/judge, a computer vision algorithm or classifier has absolutely no subjective baggages [sic], having no emotions, no biases whatsoever due to past experience, race, religion, political doctrine, gender, age, etc., no mental fatigue, no preconditioning of a bad sleep or meal. The automated inference on criminality eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together.*"

Let's look at all of this in a bit more detail.

A biased training set?

The key to understanding the problems with the Wu and Zhang paper is to look at the training sets — the images used to teach the algorithm what a non-criminal face looks like and how criminal faces differ from non-criminal ones. A machine learning algorithm can be only as good, and only as unbiased, as the training data that we provide to it.

So what did these authors provide to their algorithm as training data? They collected over 1,800 photos of Chinese men aged 18-55, with no distinguishing facial hair, scars, or tattoos.

About 1100 of these were photos of non-criminals scraped from a variety on sources on the World Wide Web using a web spider; presumably these are from professional pages of some sort because the authors know the occupation and educational background of each individual.

Just over 700 of the photos were pictures of criminals, provided by police departments. In the authors' own words, they train the algorithm using photos of

> "730 *criminals, of which 330 are published as wanted suspects....the others are provided by a city police department in China under a confidentiality agreement. We stress that the criminal face images... are normal ID photos not police. Out of the 730 criminals 235 committed violent crimes including murder, rape, assault, kidnap and robbery; the remaining 536 are convicted of non-violent crimes.*"

While ambiguous in this paragraph, the authors note elsewhere that the criminal photos are all from individuals actually convicted of crimes. Figure 2 below shows the six example photos that the authors have provided from their training set.

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 2. Criminal and non-criminal faces from Wu and Zhang (2016)

From these details alone, two massive problems leap to our attention. Each introduces major biases of precisely the sort that the authors claim are avoided by machine learning algorithms.

The first and probably most prominent source of bias in this methodology is that the images of non-criminals have been posted to websites presumably designed for promotional purposes, be they company websites or personal profiles. Many of these images will have been chosen by the photo subject himself; most of the others, while chosen by a third party, will presumably have been picked to convey a positive impression. By contrast, the images from the set of criminals are described as ID photographs. While it is unclear exactly what this means, it's a pretty good guess that these have been selected neither by the individual depicted, nor with the aim of casting that individual in a favorable light. Thank goodness no one judges our character based upon our driver's license photos!

A second source of bias is that the authors are using photographs of *convicted* criminals. As a result, even if there is some signal here, the machine algorithm could just as easily be responding to the facial features that make someone likely to be convicted by a jury, rather than the facial features correlated with actually committing a crime. Indeed it seems less plausible to us that facial features are associated with criminal tendencies than it is that they are correlated with juries' decisions to convict. We have zero prior evidence of the former claim. By contrast, a recent study (http://onlinelibrary.wiley.com/doi/10.1002/bsl.939/abstract) has demonstrated the latter. Unfortunately, it appears that unattractive individuals are more likely to be found guilty in jury trials than their more attractive peers. While the Chinese criminal system is structured differently than the US system

in which most of these studies were conducted, the judges and occasional jurors in Chinese trials may suffer from similar biases. The algorithm could be learning what sorts of facial features make one convictable, rather than criminal.

Thus while the authors claim that their algorithm is free of human biases, it may instead be picking up nothing but these biases—due to their choice of training data.

For the want of a smile

As we mentioned, the authors find that their algorithm can classify criminal faces within their data set with 90% accuracy. What facial features is it picking out that allow it to discriminate? One of the figures from their paper, reproduced below, illustrates the particular facial features that the algorithm relies upon to make the distinction between criminal and non-criminal faces.

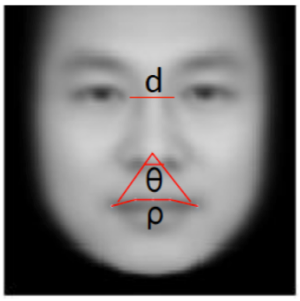

Figure 3. Facial features purportedly associated with criminality, from Wu and Zhang (2016).

The algorithm finds that criminals have shorter distances d between the inner corners of the eyes, smaller angles θ between the nose and the corners of the mouth, and higher curvature ρ to the upper lip.

Why would this possibly be?

There's a glaringly obvious explanation for the nose-mouth angle θ and the lip curvature ρ. As one smiles, the corners of the mouth spread out and the upper lip straightens. Try it yourself in the mirror.

Going back to the sample faces from the training set (our Figure 2 above), all of the criminals are frowning or scowling, while the non-criminals are faintly smiling. Now we have an alternative — and far more plausible — hypothesis for the authors findings. It is not that there are important differences in facial structure between criminals and non-criminals, it is that non-criminals are smiling in the photographs scraped from the web whereas criminals are not smiling in the photographs provided by police departments.

The authors have confused *facial features* with *facial expressions*. The former are essentially immutable aspects of facial structure, while the latter are situation-dependent configurations of contraction by the facial muscles.

The claims about detecting criminality are bullshit. All their algorithm is doing is detecting which sample set the photographs came from, based in some large part on the presence of a frown or smile.

Evaluating our interpretation

The Wu and Zhang paper provides only three criminal pictures and three non-criminal pictures from the training set, and these authors have not answered our email inquiries. Thus it hard to be certain that the six images in our Figure 2 are representative of the full training set. How then can we test our alternative hypothesis, that the algorithm is primarily classifying faces based on the presence of absence of a smile? One thing we can do is look at the composite images that the authors create. At left, composite criminal faces produced by two algorithms; at right, composite non-criminal faces produced by the same two algorithms.

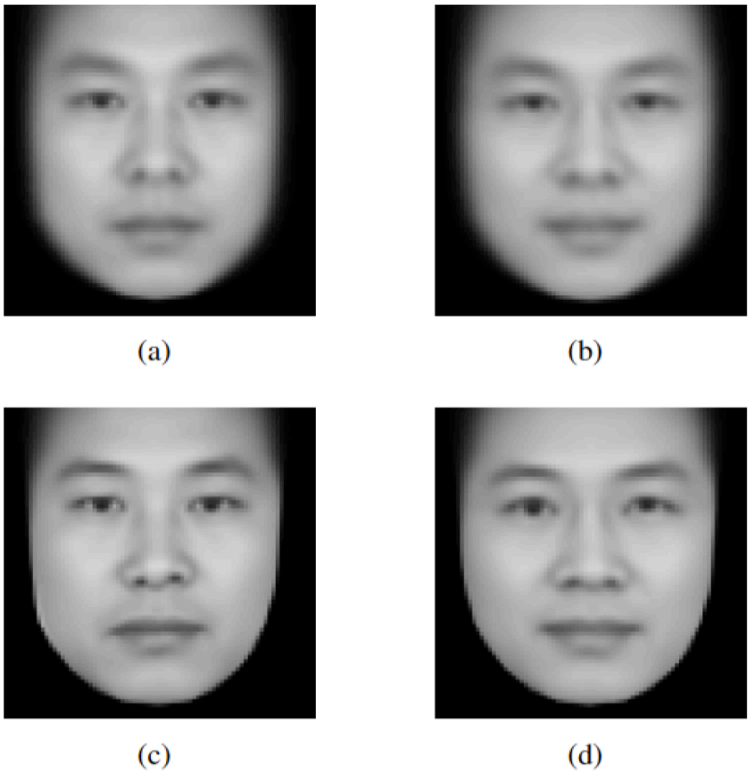

Figure 4. Composite faces for criminals (left) and non-criminals (right), as generated by two different algorithms (top and bottom). From Wu and Zhang (2016).

Looking at these faces, we cannot understand how the authors can write "*Although the antithesis of criminals and non-criminals is very strong, conventionally-defined average faces of the two populations…appear hardly distinguishable as demonstrated [in the figure above].*" Clearly the criminal composites at left are frowning, whereas the non-criminal composites at right are smiling. This strongly supports our hypothesis that the machine learning algorithm is picking up on situation-dependent facial expressions (whether a person is smiling or not) rather than underlying facial structure.

When the authors break the criminal and non-criminal down into what they call "subtypes", the smiling/non-smiling distinction becomes even more readily apparent. Below is their figure illustrating what they call four criminal subtypes at top and three non-criminal subtypes at bottom.

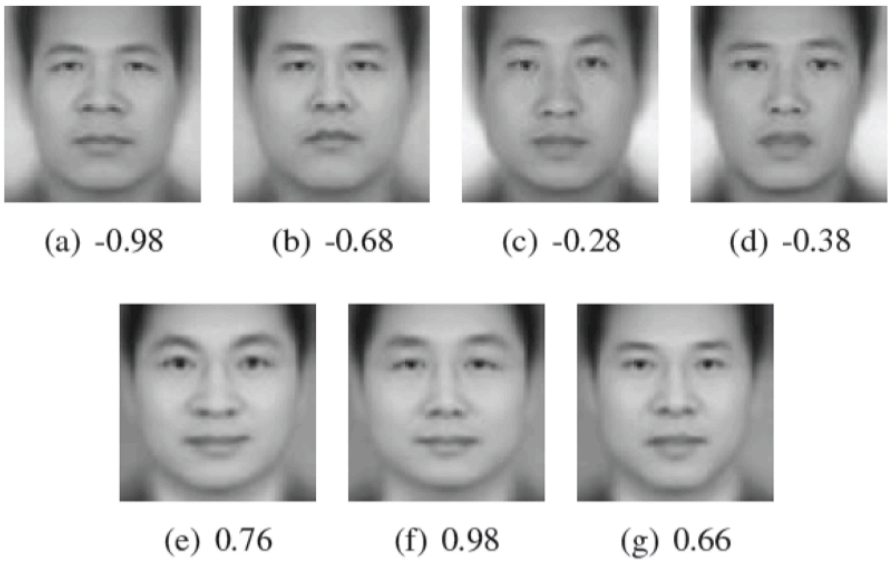

Figure 5. Purported subtypes of criminal (top) and non-criminal (bottom) faces. From Wu and Zhang (2016).

Conclusions

So where does all this leave us?

Extraordinary claims require extraordinary evidence. The authors of this paper make the extraordinary claim that facial structure reveals criminal tendencies. We have argued that, given all publicly available information, their findings can be explained by a much more reasonable hypothesis: non-criminals are more likely to be smiling in photos chosen for publicity purposes than are criminals in the ID photos (not mugshots) chosen by police departments for wanted posters and other purposes.

Notice that we did all of this without digging into the details of the machine learning algorithms at all. We didn't need to. We know that a machine learning algorithm is only as good as its training data, and we can see that the training set used here is fundamentally flawed for the purpose it is used. The implication is that one does not need technical expertise in machine learning to be able to debunk many of the bullshit claims based on such algorithms. In some cases, training data may be OK and problems may arise because of the specifics of the machine-learning algorithm, and these cases would require highly specialized knowledge to uncover. But more often, we believe, the training data will be at fault. In that case, a non-specialist can see what is going on, by thinking carefully about how a generic learning system would behave given the training data that is being used.

Doing so for this paper, we see clearly that the algorithm is not picking not up some underlying physical structures associated with criminality, but rather is discriminating based on context-specific cues from the situations under which the photographers were taking. In other words, we don't have to worry about the ethics of detecting pre-crime just yet.

Authors' response

We reached out the authors of the study, Xiaolin Wu and Xi Zhang, and offered them an opportunity to respond to this case study. They kindly sent a detailed letter, which with their permission we have posted below.

In the first main paragraph of their response, they pose an interesting alternative to our smile hypothesis: that the algorithm is picking up facial relaxation, rather than the presence or absence of smile *per se*. We are particularly intrigued by their note that perception of the facial expressions may be in part culturally dependent.

Much or all of the remainder appears to be a generic response to critics; it addresses several points that we did not make in our article, such as overfitting (we don't think this is the issue) and the white collars (which we assumed, correctly, it seems, that they had masked out). Readers may decide for themselves whether their precautions and caveats are sufficient, given the sensitivity of the subject matter and potential for misuse of their methods.

Xiaolin Wu and Xi Zhang write:

We welcome sober and fair academic discussions, instead of name calling, surrounding our paper.

In our experiments, we did control facial expressions, such as smile and sad, but not faint micro-expressions (e.g., relaxed vs. strained). We intend to exert much tighter control on facial micro-expressions in the future as soon as a reliable algorithm reaches the sophistication to do so. Regarding to the face "subtypes" found in our paper, we and our students and colleagues do not think the difference between the top and bottom rows of Figure 13 is smiling versus non-smiling; instead, the faces in the bottom row appear seemingly more relaxed than those in the top row. Perhaps, the different perceptions here are due to culture difference.

All criminal ID photos are government issued, but not mug shots; to our best knowledge, they are normal government issued photos like those for driver's license in USA. In contrast, most of the noncriminal ID style photos are taken officially by some organizations (such as real estate companies, law firms, etc.) for their websites. We stress that they are not selfies.

Our critics are quick to point out the relatively small sample set used in our experiments and the risk of data overfitting. We are sorely aware of this weakness but cannot get more ID images of convicted Chinese males for obvious reasons (the ongoing publicity might have dashed all our hopes to enrich our data set). However, we did make our best efforts to validate our findings in Section 3.3 of our paper, which opened as follows but completely ignored by our critics.

> *"Given the high social sensitivities and repercussions of our topic and skeptics on physiognomy [19], we try to excise maximum caution before publishing our results. In playing devil's advocate, we design and conduct the following experiments to challenge the validity of the tested classifiers ..."*

We randomly label the faces of our training set as negative and positive instances with equal probability, and run all four classifiers to test if any of them can separate the randomly labeled face images with a chance better than flipping a coin. All face classifiers fail the above test and other similar, more challenging tests (refer to our paper for details). These empirical findings suggest that the good classification performances reported in our paper are not due to data overfitting; otherwise, given the same size and type of sample set, the classifiers would also be able to separate randomly labeled data.

Regarding to the wearing of white-collared shirts by some men but not by others in the ID portraits used in our experiments, we did segment the face portion out of all ID images. The face-only images are used in training and testing. The complete ID portraits are presented in our paper only for illustration purposes. We did not spell out this data preparation detail because it is a standard practice in the field of machine learning.

Nevertheless, the cue of white collar exposes an important detail that we owe the readers an apology. That is, we could not control for socioeconomic status of the gentlemen whose ID photos were used in our experiments. Not because we did not want to, but we did not have access to the metadata due to confidentiality issues. Now reflecting on this nuance, we speculate that the performance of our face classifiers would drop if the image data were controlled for socioeconomic status. Immediately a corollary of social injustice might follow, we suppose. In fact, this is precisely why we said our results might have significance to social sciences.

In our paper, we have also taken steps to prevent the machine learning methods, CNN in particular, from picking up superficial differences between images, such as compression noises and different cameras (Section 3.3).

It should be abundantly clear, for anyone who reads our paper with a neutral mind setting, that our only motive is to know if machine learning has the potential of acquiring humanlike social perceptions of faces, despite the complexity and subtlety of such perceptions that are functions of both the observed and the observer. Our inquiry is to push the envelope and extend the research on automated face recognition from the biometric dimension (e.g., determining the

race, gender, age, facial expression, etc.) to the sociopsychological dimension. We are merely interested in the distinct possibility of teaching machines to pass the Turing test on the task of duplicating humans in their first impressions (e.g., personality traits, mannerism, demeanor, etc.) of a stranger. The face perception of criminality was expediently (unfortunately to us in hindsight) chosen as an easy test case, at least in our intuition as explained in our paper:

> *"For validating the hypothesis on the correlations between the innate traits and social behaviors of a person and the physical characteristics of that person's face, it would be hard pushed to find a more convincing experiment than examining the success rates of discriminating between criminals and non-criminals with modern automatic classifiers. These two populations should be among the easiest to differentiate, if social attributes and facial features are correlated, because being a criminal requires a host of abnormal (outlier) personal traits. If the classification rate turns out low, then the validity of face-induced social inference can be safely negated."*

We agree that the pungent word criminality should be put in quotation marks; a caveat about the possible biases in the input data should be issued. Taking a court conviction at its face value, i.e., as the "ground truth" for machine learning, was indeed a serious oversight on our part.

— Xiaolin Wu and Xi Zhang

Calling Bullshit has been developed by Carl Bergstrom (http://ctbergstrom.com) and Jevin West (http://www.jevinwest.org) to meet what we see as a major need in higher education nationwide.

(https://twitter.com/callin_bull)

(https://www.facebook.com/callinBS/)

(mailto:bullsht.course@gmail.com)