Implementation of Artificial Surveillance in Schools

Bill Shao

September 2020

Contents

1 Abstract

The quick rise and implementation of machine learning has led to its application in many fields such as advertising, security, and marketing. During this time, it is important to consider potential bias and public privacy when it comes to surveillance technology, and the potential overreach an algorithm may have when unregulated. This study considers both the factors of algorithmic bias and policy implementation on the topic of facial recognition surveillance in school campuses, and proposes both technological and policy level improvements to this growing problem. These proposals are discussed in three major topic areas, racial profiling, surveillance decisions, and mission creeping. Ultimately, this study concludes that machine learning, if used at all in schools, should first be supported by large policy changes to data-handling within schools and also refined for bias to prevent amplification of racial divisions.

2 Introduction

This paper proposes solutions to the growing concern over the use of facial recognition technology to bolster school security. In recent years, the advancement of facial recognition technology has led to many implementations in public use as a potential security system. At first, this technology was utilized in phones as well as public spaces such as airports or public parks. From there, it expanded to other uses, schools being among them [9]. Motivation for this expansion comes from recent news regarding dangerous intrusions on school spaces, such as ones conducted by armed shooters and sex offenders [21]. Facial recognition is seen as a viable solution to these problems due to their ability to quickly cross-reference dangerous individuals to large datasets and simultaneously cover large areas of campuses. Proposals and trial implementations recently have led to a large amount of the backlash, primarily against the intrusion of privacy, data usage concerns, and racial bias [17]. This report considers the validity of these three complaints, as well as potential solutions to them.

(1) Racial Profiling: Across all applications, facial recognition has been cited to identity minorities correctly at lower rates [5]. This has led to the idea that the technology could be racist, and discriminate against minorities. In schools specifically, many are concerned that this problem will exacerbate the school-prison pipeline by misidentifying students of color at greater rates, leading to more incorrect detentions and hazardous scenarios [11]. For example, greater amounts of false alarms or identifications may lead to more conflict-prone scenarios occurring between police and students of color, which would result in worse treatment and targeting of those students. This study goes into potential technical solutions to this problem, as well as how it would impact the current school-prison pipeline.

(2) Surveillance Decisions: When considering the use of surveillance, the ques-

tion of camera placement, decision confidence, etc. must be noted as potentially affecting their efficacy. For example, utilizing cameras only at the entrance of a school may leave other areas vulnerable to intrusion, but driving up the usage of cameras will also lead to a greater risk in misidentifying students due to higher chances to incorrectly assess someone. This section explores the balancing aspects of facial recognition usage.

(3) Mission Creeping: Another larger consideration regarding facial recognition is the issue of data-harvesting, which is also known as mission creeping, where the use of data extends beyond its intended purpose. Many against the implementation of facial recognition in schools cite this reason as to why it shouldn't be used, claiming it violates right to privacy and other related issues [17] [11]. This concern is exacerbated by machine learning data particularly, since the various patterns and features it detects can be applied to unrelated usages. Since there is no technical fix for this, this section focuses on policy approaches to the problem.

3 Related Work

3.1 Identification Objective

Currently, there are many growing solutions that stray away from the issue of racial bias in facial recognition systems by opting for object detection instead. These systems focus around guns and harmful weapons, since other intruders such as sex offenders are not identifiable by carried objects [14] [1]. Although some potential for bias still exists around object detection, this is most likely mitigated due to the fact that the application area is consistent across the US and not enough data exists discussing the pattern [8]. Still, this approach doesn't address sex offenders and other dangerous people on school campuses, so this study will directly address this under-covered issue.

3.2 Bias Prevention

On the topic of Bias Prevention in Machine Learning, two main approaches have been pursued, being dataset modification and algorithmic solutions. Examples of algorithmic solutions include Domain Discriminative Training and Domain Adversarial Training, which try to be race-blind and race-conscious when training respectively [22]. The other approach, which this paper will cover, is through datasets, which has been done in the past by datasets such as FairFace and VGGFace, which create datasets with attention to proportions of minorities and genders [15][6]. The general idea of these approaches is to create equally proportioned datasets so that, when training, each race appears equally and the model can train for each the same amount.

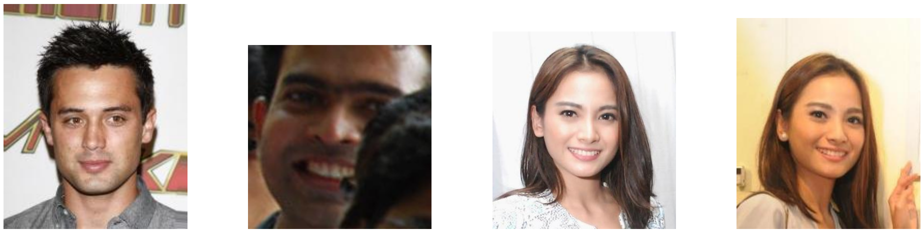
3.3 Policy Action

Another important area of existing study are approaches to policy regulations and control on the use of machine learning. Currently, this topic is being heavily pursued in fields such as Healthcare and Policing, but is relatively underdeveloped in the school environment due to the growing but lesser implementation of it in education [4]. Still, many observations can be cross-applied, such as the issue in police use of manipulating prediction results to get skewed results [4]. Another question is how bias will interact with non-discriminatory policies in schools, since existing laws that define biased practices may not directly apply to AI [12]. Also, since existing policies like the Family Educational Rights and Privacy Act don't cover the usage of data within schools, which can become troublesome with AI due to the potential for unintended use cases [4]. This paper will address some unique flaws or gray areas in existing policy with regards to AI, and propose solutions or directions of approach.

4 Methods

4.1 Dataset Selection

This study utilizes three popular and commonly used face datasets in order to compose the resultant dataset. These datasets were CelebA, VGG Face, and FairFace. When considering datasets, size, available labels, and data composition were all considered as influencing factors. Dataset size was a factor primarily based on technical capability, as a dataset of too small size would either be out-represented by the others or make it so that training set itself is too small to properly train on. All three of the above selected were considerably sized, meaning the dataset generated from them would be adequate [18] [6] [15]. A secondary issue raised is the available labels for the dataset. Since the datasets are combined, there must be at least one unifying label that can be trained on. Ultimately, both the CelebA dataset and FairFace dataset shared



(a) CelebA Example.

(b) FairFace Example.

(c) VGG Example.

(d) Corresponding Person VGG Example.

Figure 1: Sample Images from Compiled Datasets.

the label of gender, and the VGG-Face dataset was grouped based on multiple photos of a singular person, making hand-labeling a viable option. Lastly, the data composition or type-of-data was considered in this task in order to ensure a wide variety of data sources as a preventable measure for bias in this task. CelebA is composed of online sourced images of celebrities, meaning the photos are unique to them (Fig. 1a) [18]. Likewise, VGG-face is composed of folders of multiple photos per individual in dataset (Fig. 1c & 1d) [6]. Lastly, FairFace considers more pedestrian, commonplace faces, also differentiating itself from the others (Fig. 1b) [15].

Summary: In order to create the dataset, various factors were considered when selecting datasets to build from. These factors helped to ensure that the dataset would work technically, as well as satisfy the purpose of testing a new method to reduce bias. Ultimately, the datasets that fulfilled these requirements were the CelebA, VGG-Face, and FairFace datasets.

4.2 Dataset Compilation

As hand-labeling was required for the VGG dataset, overall dataset size was set at 100000, which allotted enough images to create a considerable sized training and validation set of 85000 and 12500 respectively. Equal representation of each dataset was favored as to not over or under-represent each other, but due to the image grouping of the VGG dataset, values were not fully equal between the three datasets. The dataset was composed of 35000 randomly selected images from the CelebA and FairFace datasets, as well as 32000 selected images from the VGG dataset. The dataset was randomly shuffled and normalized prior to training.

4.3 Model Selection and Training

The model utilized for diagnosing the dataset is a pre-trained Resnet-18 model with an altered FC section of a singular linear layer of (512,1), which predicts gender. The feature section remains untouched, and only the classifier was changed. Resnet was chosen due to it being a standard, modern CNN that could be easily adapted to the use case of identifying facial characteristics. Furthermore, since it doesn't impose any form of bias mitigation or reduction itself, the datasets could be more thoroughly tested through this model.

Binary Cross-Entropy Loss was used for the loss function of this model, and a Sigmoid function was used to normalize the outputs of the model. The optimizer used for this model was Stochastic Gradient Descent, and it employed a learning rate of 0.00035 and momentum of 0.8 for the benchmark test, and a learning rate of 0.00025 and momentum of 0.8 for the new dataset. Batches were created of size 64, and 10 epochs were run for both models. Each image was cropped to size 256 and normalized to standard Resnet-18 parameters. Accuracy, Positive Parity Value, and Negative Parity Value was tracked at each epoch, and raw sigmoid outputs were also stored in order to conduct confidence testing.

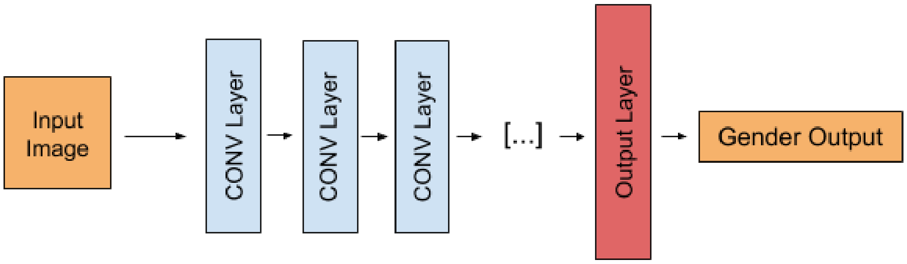


Figure 2: Basic diagram of Resnet-15 Model Used.

Summary: A standard, vanilla machine learning model structure (Fig. 2) for image recognition was used to test the dataset. Although models with bias reduction exist, in order to accurately gauge the affect of the dataset change alone, one without such features was selected. The structure was taught to predict gender, and parameters were tuned in order to maximize recorded result statistics.

5 Results

Dataset	Accuracy	PPV	NPV	Parity Difference
Benchmark (CelebA)	59.29%	0.586	0.700	0.113
Testset (VGG, FF, CelebA)	76.91%	0.753	0.790	0.037

$p(y|t) = \#$ of Images with predicted label y and true label t.

$$PPV = \frac{p(y=1|t=1)}{p(y=1|t=1)+p(y=1|t=0)}$$
$$NPV = \frac{p(y=0|t=0)}{p(y=0|t=0)+p(y=0|t=1)}$$

[10]

6 Discussion

6.1 Racial Profiling

When considering improvement on racial bias, accuracy is often not a clear or useful determiner. This is due to the fact that accuracy makes it hard to distinguish predictive trends in the data by categorizing all the data into incorrect or correct prediction categories [10][7]. Furthermore, in most datasets, an unbalanced dataset will make it so that, although the results may have high accuracy, the minority prediction rate is still low but less significant since they make up less of the total [7]. Instead, the Positive Parity Value (PPV) and Negative Parity Value (NPV) were collected as result statistics in order to assess the data. In this case, the PPV represents the model's ability to predict males correctly while the NPV represents the ability to predict females correctly. In both these categories, it is clear that the Testset performs better than just the benchmark of CelebA with a PPV of 0.753 compared to 0.583 and a NPV of 0.79 compared to 0.7. Furthermore, the Parity Difference is also less, meaning that the Testset-trained model has relatively equal ability to predict men and women, making less biased.

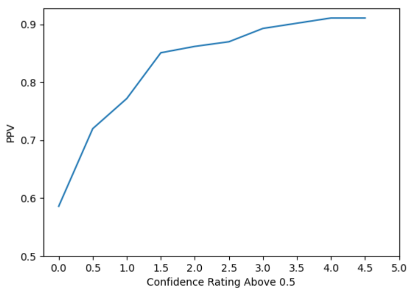
The comparison of these results to existing data and patterns of discrimination in the school-prison pipeline is more challenging. Since technical limitations make it impossible to get direct 1-to-1 mapping for the facial recognition program and because little data exists about facial recognition discrimination in trial cases, the most probably comparison in a qualitative look at known statistics today and patterns in the data results.

Since this system exists in order to determine potential threats to on school safety, the analysis is concentrated on the % difference of school based arrests and encounters with law enforcement rather than disciplinary actions such as suspensions and expulsions, because they would be most likely to occur in dangerous environments [23]. Currently, there is a drastic difference between the percentage of students of color who are arrested compared to their white peers [23]. Around 70% of all arrested or referred students were Black or Hispanic during the 2009 and 2014 school year [13] [23]. In comparison to the parity difference of the testset, the testset shows a much smaller variation, but this may only be due to the fact that it has to predict gender, not map faces to people. Still, the ongoing drastic bias of school arrests does have the potential of being alleviated through the replacement of traditional law enforcement with facial recognition.

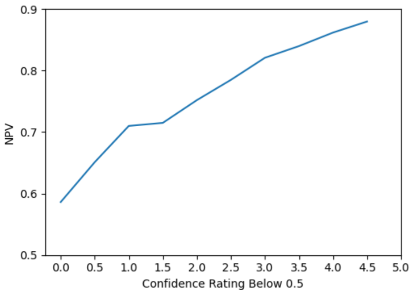
6.2 Surveillance Decisions

Another consideration in the use of facial recognition technology is the confidence rating and location of cameras located around a school. In response to many concerns over in-class and hallway monitoring, one popular proposition is that cameras only be used in outdoors, major-traffic areas of the school to detect entrances and exits from the school, but not indoors activity [20][9]. This solution is most likely preferable, but the impact of having vs. not having indoors monitoring has yet to be assessed.

Confidence rating is based on how confident an algorithm needs to be in order to label someone. In the results taken above, there was no confidence rating and anything below 0.5 was rounded down while anything above 0.5 was rounded up. Including a confidence rating will reduce the chance of having a false positive due to low confidence but also decrease the overall number of predictions, allowing people to ”slip-through” the detection system [19].



(a) PPV Confidence Graph



(b) NPV Confidence Graph

By assessing the results and dropping out values that don’t meet various confidence thresholds, a graph can be generated to compare the change in Parity Value when confidence is increased or decreased below the threshold. The most notable difference between the two graphs is shape, as the PPV graph follows a logarithmic shape while the NPV graph is mostly linear. This indicates that the negative label is more biased, since it means there is considerable misidentification at higher confidence levels in comparison to the PPV graph. Another note is that the PPV graph levels off around a confidence rating of 1.5-2, meaning it is most optimal to select a point there in order to still consider cases while decreasing the chance of a false alarm.

6.3 Mission Creeping

A final major concern that lies in facial recognition is right to privacy and data usage. This problem has very few technical solutions since data collection will always be necessary to identify faces, and small scale solutions such as ones discussed in Section 6.2 only provide minimal fixes. Thus, policy enactment is key to resolving this problem. Currently, two main policy laws exist to protect student and children data, the Family Educational Rights and Privacy Act (FERPA), and the Children's Online Privacy Protection Act (COPPA) [2] [16]. These laws mostly serve to regulate the handling of student data to parents and other entities, and have mostly been abided by, despite recent technological booms in classrooms [2]. The focus around surveillance, however, is the use of data within the school and not just with outside entities [17] [21] [3]. On this topic, many trial schools and boards are still drafting regulation for the technology [21]. Although some advocate for its usage in emotion / engagement monitoring and more personal aspects of student life, its application in such fields is indecisive and negatively perceived [3][9]. Thus, current regulation should limit the usage of facial recognition data to surveillance only, and ban it from other applications.

Something of note is that current school-policy interaction on this topic is shaky, as in the case of the Lockport school district Facial Recognition was hastily implemented before prior policy was rolled out, meaning that stronger enforcement or regulation of future potential changes should be present to ensure it is abided by [21].

6.4 Comparison of Pro-Cons

A wide-scale, open-ended implementation of facial recognition in schools will almost never occur due to backlash as well as wide ranging privacy and social climate concerns voiced over its usage [3][9][21]. As such, facial recognition, if used at all in school environments, should be heavily limited to its specific use case. Noted previously, the technology does provide some potential improvement in its ability to reduce bias in detecting people if replacing the current law enforcement-school interaction. This could potentially lead to less biased incarcerations or dangerous events, reducing the amount of instances where students of color are disadvantaged [23]. This value can also be further fine tuned through the usage of confidence thresholds, and individual changes can be made on a per school basis regarding camera placement or count to minimize intrusion. Still, the topic of privacy concerns remains the biggest drawback, since little currently exists and the precedent for such programs to abide by such policies is unclear [21]. Although legitimate trials with greater data collection should be done in order to assess the numerical efficacy of its benefits, if substantial groundwork and policies are provided to help control facial recognition, it could prove a useful tool for tracking dangerous individuals in school campuses while also reducing racist discipline.

References

[1]

[2] Family educational rights and privacy act (ferpa), Mar 2018.

[3] Mark Andrejevic and Neil Selwyn. Facial recognition technology in schools: critical questions and concerns. *Learning, Media and Technology*, 45(2):115–128, 2020.

[4] Ben Buchanan and Taylor Miller. Machine learning for policymakers., 2017.

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

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

[7] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *CoRR*, abs/1810.08810, 2018.

[8] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? *CoRR*, abs/1906.02659, 2019.

[9] Jeremy Engle. Should facial recognition technology be used in schools?, Feb 2020.

[10] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis, 2020.

[11] Vaidya Gullapalli, Vanessa A. Bee, and Sarah Lustbader. New to the school-to-prison pipeline: Armed teachers, facial recognition, and first-graders labeled 'high-level' threats, 2019.

[12] Dr. Bethanie Hansen. Artificial intelligence in education: Where are the laws?, Mar 2020.

[13] Evie Blad Harwin and Alex. Analysis reveals racial disparities in school arrests, Feb 2017.

[14] Jeremy Hsu. Ai aims to save kids from shooters, Aug 2018.

[15] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age, 2019.

[16] LearnPlatform. Student data privacy regulations across the us: A look at how california, illinois, and new york are handling privacy and what it means for k-12 leas, Mar 2020.

[17] Mark Lieberman. Facial recognition tech in schools prompts lawsuit, renewed racial bias concerns, Jun 2020.

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[19] Amit Mandelbaum and Daphna Weinshall. Distance-based confidence score for neural network classifiers. *CoRR*, abs/1709.09844, 2017.

[20] Alfred Ng. Even facial recognition supporters say the tech won't stop school shootings, 2020.

[21] Thomas J. Prohaska. Lockport schools testing facial recognition despite state warning, Jun 2019.

[22] Zeyu Wang, Klint Qinami, Yannis Karakozis, Kyle Genova, P. Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8916–8925, 2020.

[23] Rachel Wilf. Disparities in school discipline move students of color toward prison, Mar 2012.