An investigation into Bias in Facial Recognition using Learning Algorithms

MSc Research Project
Data Analytics

Annye Braca
X14121212

School of Computing
National College of Ireland

Supervisor: Dr.Simon Caton

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Annye Braca
Student ID:	X14121212
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Simon Caton
Submission Due Date:	11/12/2017
Project Title:	An investigation into bias in facial recognition using learning algorithms
Word Count:	XXX

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

An investigation into Bias in Facial Recognition using Learning Algorithms

Annye Braca
X14121212
MSc Research Project in Data Analytics

11th December 2017

Abstract

Data Science has become a field of major activity in recent years. Whether it involves a clients propensity to churn, making medical diagnoses or inferring dating tendencies, the areas of classification and prediction are inextricably linked to this field. This research project involved the investigation of the potential to apply artificial intelligence as a discriminator of criminal tendencies in people. In addition to creating a model that attempts to classify criminals and non-criminal images, the research was also concerned with the potential biases that may exist within a model of this nature. A deep learning algorithm was utilised for the classification problem and applied to numerous image subsets with consideration given to gender and emotion. Once converted to arrays, the images were high dimensional and principal components analysis was conducted as a dimension reduction technique. Results included stratified 10-fold cross validation accuracies, confusion matrices, learning curves and emotion histograms.

1 Introduction

Face image analysis is a research field that aims to identify personality traits by analysing peoples faces using automated facial features that map those features to teach the algorithm to identify patterns and classify them. Our faces might disclose more than we believe. Machine Learning attempts to analyse the face of a person and infer certain characteristics about them. One recent application has involved attempting to identify whether individuals faces betray criminal tendencies. Algorithms have shown ability to discriminate more accurately than humans do. Machine learning algorithms can look at a persons face and envision character and traits that are undetectable to the human eye. The danger of this technology lies in its imperfections. The challenge is to improve the effectiveness of algorithms and reduce the effect of biases in their analysis. Artificial intelligence systems face many challenges. How do we evaluate a classifier that tries to flag specific face attributes such as a tendency towards terrorism or paedophilia. This could lead to many false positives. When can we say a machine learning model is actually good? The aim of this paper is to investigate the presence and effects of biases in learning algorithms by focusing on the facial recognition features pattern identification problem for computer vision and machine learning technology tasks. The work presented by these paper was built on a Python framework composed by OpenCV [1]for feature extraction, Theano(Al-Rfou et al.; 2016) for fast computation of mathematical expressions, Keras[2] Neural Network API then Scikit-Learn[3]workflow for training the model and finding its accuracy. The paper first presents a baseline outcome for criminality recognition with bias issues being examined thereafter. Also, the learning algorithms are tested on different corpus of test data, that aims to understand the existence of biases and their impact on the recognition rate. The paper concluded by discussing the problem of bias in an extensive training set as future work

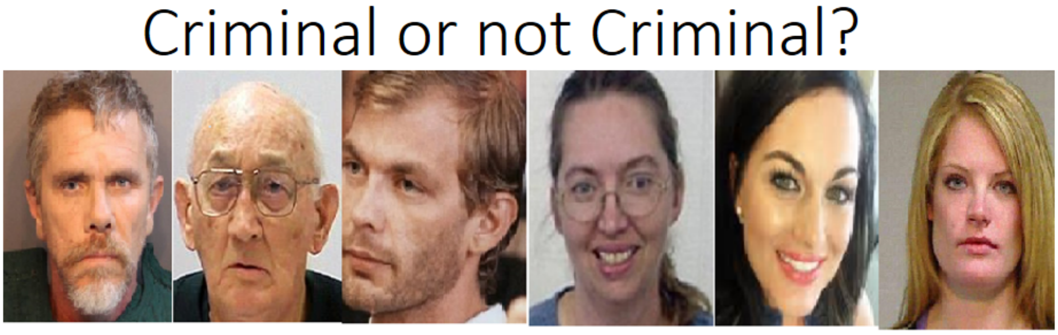

Figure 1: Image-set sample

2 Related Work

The face is one of the most recognisable elements of a human being possessing many differentiators. The face is most likely the primary means of recognition of a person, transmitting information, communicating with others, inferring peoples feeling, and so forth. From face images, information can inform on many traits, such as race, gender, age, health, emotion, psychology, profession and many others. There is proof of the links between personality traits and facial features. (Zebrowitz; 1998).) Through time many researchers have been trying to validate the hypothesis that peoples character can be judged from their faces. Cesare Lombroso[4] in 1871 was the first to realize that crime and criminals could be studied scientifically. He postulated that thieves could be identified by their facial expressions. Recently, further work by (Wu and Zhang; 2016)has revisited this theory, motivated by advances in the computer vision field and AI. They aimed to demonstrate the correlation between criminality and features of the human face. They suggest that their software can identify a criminal face with an accuracy of 90% with the claim their research is free of biases. The study has elements of the historical area of phrenology. The hypothesis is that the facial structure reveals criminal tendencies. They intended to teach the algorithm to distinguish what a criminal face looks like and how to learn the difference between both classes (criminal and non-criminal). They trained

[1] https://opencv.org/

[2] https://keras.io//

[3] http://scikit-learn.org/stable/

[4] http://www.historyextra.com/article/feature/born-criminal-lombroso-origins-modern-criminology

four classifiers (logistic regression, KNN, SVM, CNN) with data that was collected from the internet and controlled by gender, race and facial expression A problem is that most data extracted from the internet can be biased. Non-criminal images may show a positive looking impression whereas criminals often tend to show an angry or sad expression which may be a natural reaction to their circumstances. The algorithm could be classifying facial emotions as opposed to some kind of latent criminality. An additional source of bias may be an inherent bias within the judicial system. Many criminals are convicted by jurys and there is evidence to suggest that people have a preconception of what a criminal looks like. The earliest paper published about how computer vision may correlate to social attributes and facial features was published by (Wang and Kosinski; 2017).The researchers adopted a deep neural network to extract features and classify peoples sexual orientation given a single facial image. In contrast with Lombroso and the be born criminal theory, Wang and Kosinski aim was to demonstrate using Machine Learning the prenatal hormone theory of sexuality. They relied on Gaussian blur for obtaining the average probability of being gay. Masking a given area of the image, the most informative areas are the nose, eyes, eyebrows, chin - those facial points showed to be gender atypical. The major bias of this research is the fact that researchers studied just images of white people, which constrains the study and prevents it from being universally applicable. An extended face analysis research using Machine learning was made by (Reece and Danforth; 2016)They used Instagram face images to detect markers of depressions and early psychiatric disorders. They used colour analysis, metadata components and face detection algorithms. This research used an ensemble of computational methods from machine learning and image processing of psychology markers which were successfully applied to identify and predict depression in Instagram users.(X.Geng and Yin; 2013) presented a computer-based age estimation model. Significant progress was achieved using supervised learning; their work consisted of extracting images features and a learner age estimator. Deep learning methods are increasing in popularity in the field.(A. and Sutskever; n.d.)) presented a scheme for image classification with deep learning. Furthermore with the advance of computer vision researchers are extending facial recognition to facial expressions and micro-expression (Ekman; 1978). The goal is to teach the algorithm to recognize the six universal emotional facial expressions: Happiness, Surprise, Sadness, Disgust, Anger, Fear. The first step for facial expression analysis is face detection followed by facial features extractions such as eyes, eyebrows and mouthIn the last decade many papers focusing on human emotions has been published.(Tsapatsoulis; 2000) used fuzzy inference for emotion classification whereas(Oliver; n.d.) proposed a Hidden Markov Models for facial expression recognition based on the real-time tracking of mouth shape.(Zhang; 2005)use Bayesian network classifier to recognize the six basic facial expressions also combined Facial and voice analysis.(Kutne et al.; 2001)chose to use Mahalanobis distance and measure each expression as a vector.(Wang et al.; 2017) studied PCA[5] to determine the principal component pairs to estimate the degree of emotion from facial expressions (for example less Happy, moderately happy, and very happy)

2.1 Face Detection and Recognition

Face detection algorithms are composed of many methods and techniques for localizing and extracting face regions from the rest of the image. There are many techniques. One of the classical techniques is the use of a weak classifier cascade for face detection. These

[5] https://en.wikipedia.org/wiki/Principal$_{c}omponent_{a}nalysis$

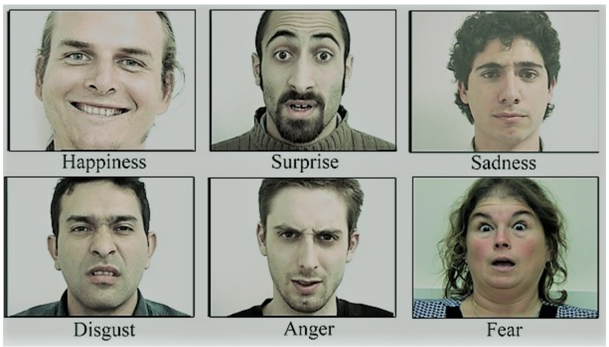

Figure 2: Universal Human Emotions

techniques were proposed by (Viola and Jones; 2001).OpenCV provides Haar classifier in the AdaBoost cascade, Adaboost algorithm (Zheng; 2005) implements different mechanisms for fast and efficient extraction of features, HFFD is a feature-based method which used AdaBoost and cascade classifier named lbpcascadefrontalface for frontal face detection. for Feature Selection Haar cascades select just the important features from a large number of features like features representing eye area, nose area and mouth area of a human face, it eliminates a significant amount of unnecessary computation during the training process.

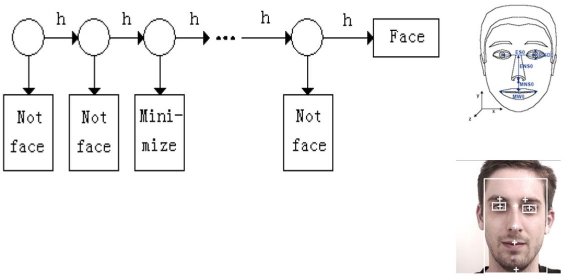

Figure 3: Face Detection and Recognition

As shown above the process involves four main steps:

1. Haar Feature Selection:All human faces share similar characteristics such eyes region being darker than the nose bridge region These are known as digital image features based upon Haar basis functions.(M.-H. Yang and Ahuja.; 2002)

2. Creating an Integral Image:Through computing the rectangles adjacent to the rectangle present at (x,y)a single image representation can be achieved which helps in speeding the procedure.

3. Adaboost Training:This algorithm finds small critical visual features from a large set of potential features.

4. Cascades Classifiers:Process of combining the classifiers performed on face-like regions.

2.2 Image Processing

Principal Components Analysis (PCA) based methods are widely used for image face recognition and processing. The main idea of one-dimensional PCA for face recognition is to get an eigenspace projection. These methods rely on two points. Firstly, the pattern of similarity of the observations and the variables can be represented as points on maps by PCA (Turk and Pentland; n.d.). Secondly, the similarity of face images is evaluated by calculating the distance between these points.(Saporta and Niang; 1996) On the other hand, when used to reduce dimensions on a digital image the method extracts the principal pattern of a linear system. The dimension reduction process by PCA generally consists of four major steps: (1) Normalize image data (2) Calculate covariance matrix from the image data (3) Perform Singulr Value Decomposition (SVD)(K. Dabov and Egiazarian; 2007) PCA processes the less significant features of the image data by locating the greatest variance on distributed datasets. These principle components hold the feature factor information of the original image; afterwards the image still maintains its principal characteristics(Wang et al.; 2017)

3 Methodology

Data compilation for both data sets (criminal and non-criminal) was one of the challenges of this paper. To build the criminal data-set, images were scraped from the Web using the BeautifulSoup [6] package in Python[7].Sources included the FBI[8] wanted-people public databases, and US police mugshot repositories[9] More than 1000 criminals of mixed race, gender, and crime type were collected. The non-criminal data-set was scraped from Google-images [10] and various face databases[11] with images being of mixed gender, race and expression. The classification framework involved using OpenCV to preprocess images, PCA for dimension reduction and a Keras Sequential Deep Learning Neural Net for binary classification (criminal and non-criminal). A second framework performed emotion classification using OpenCVs FischerFaceRecognizer.

4 Implementation

The model begins by reading in approximately 2000 images of criminals and non-criminals. It utilises various OpenCv operations such as image readers, grayscale converters and face detectors. The model will reject any images that are not successfully read and processed by OpenCv. The surviving images will have been cropped, converted to grayscale, aligned and vectorised. Each image is then on observation with 40,000 features (200 x 200 pixels). In order to reduce dimensions, PCA is applied to the design matrix with a view to keeping the components that explain the most variance in the data. A graph of number of components vs. explained variance was used to optimise the choice of number of principal components. Once the dimension reduced design matrix was achieved, this, along with the target vector, was fed to a Keras wrapped Sequential Deep Learning neural network

[6] https://pypi.python.org/pypi/beautifulsoup4

[7] https://www.python.org/

[8] https://www.fbi.gov/wanted

[9] http://mugshots.com/US-Counties/

[10] https://images.google.com/

[11] http://www.face-rec.org/databases/

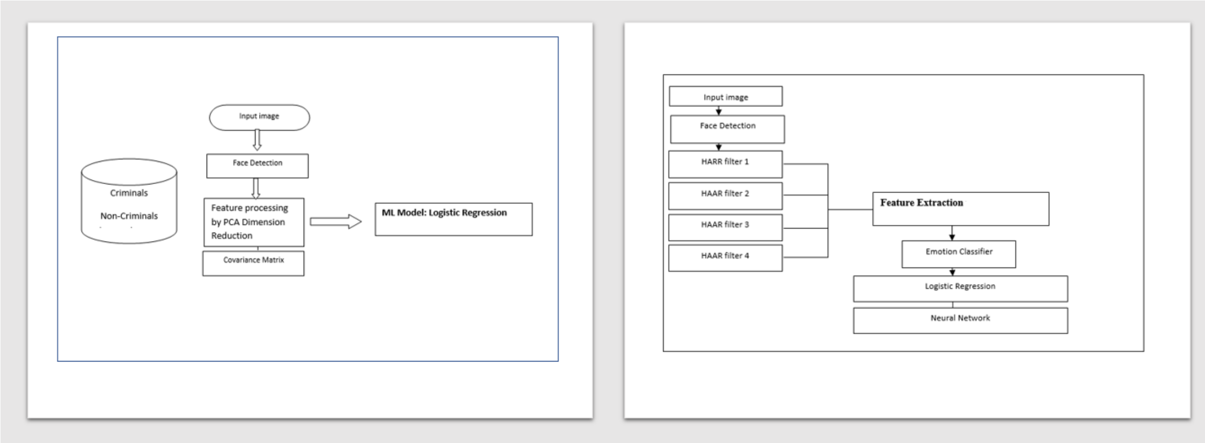

Figure 4: Framework

for binary classification. The Neural network architecture was tuned by varying the width and depth of the network (adding Dense layers and varying the number of neurons) and by adjusting hyperparameters such as optimser, learning rate and activation functions. Dropout was also applied to help prevent co-adaptation of neurons and over-fitting. Performance metrics were a stratified 10-fold cross validation accuracy measure, confusion matrices and learning curves. Class imbalance in the data was minimal with the most extreme case being 44% non-criminal an 56% criminal. Three cases were investigated for classification efficacy. The first used the complete image set of 621 criminals and 481 non-criminals. These contained mixed gender, race and wide age spans. The second set used was a set of approximately 300 women (non-criminal and 300 men criminal. The idea was to see if the classifier could achieve a better accuracy given the gender differentiator. Lastly, a set of approximately 100 women (criminal) and approximately 100 men (non-criminal) were investigated. As an additional line of research, an emotion classification model was built to attempt to classify the emotional profiles of the two image sets (criminal and non-criminal). The model primarily leveraged OpenCvs FischerFaceRecognizer. The Cohn-Kanade image set composed of approximately 100 posers was used to train and test the emotion classifiers. The model was then used to create histographic plots, across both image sets, of the frequency of the 8 core emotions ((0)neutral, (1)anger, (2)contempt, (3)disgust, (4)fear, (5)happy, (6)sadness, (7)surprise.

Figure 5: Sample Criminals faces

5 Evaluation

5.1 Results Set 1 All Images

(Mixed gender, race and emotion)

- Criminal images: 621
- Non-criminal images: 481
- No. of Principal Components: 750
- Explained Variance: 99%
- Stratified 10-Fold Cross Validation Accuracy: 60.8%

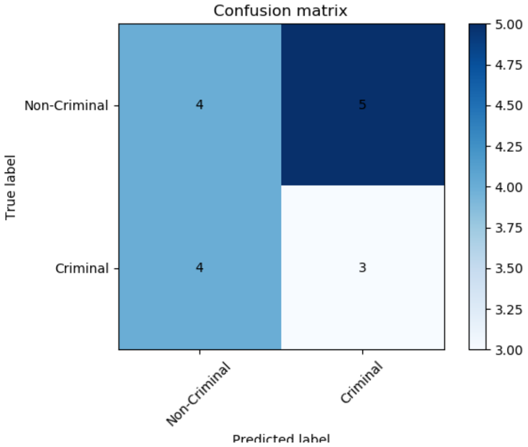

Figure 6: Confusion Matrix All Images

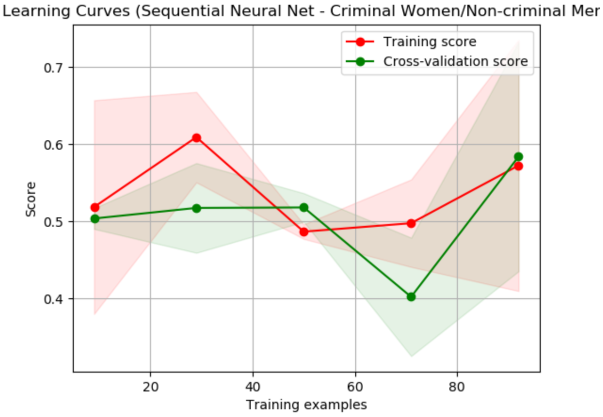

Figure 7: Learning Curve - All Images

5.2 Results Set 2 Criminal Men vs. Non-criminal women

(Mixed race and emotion)

- Criminal images (Men): 240
- Non-criminal images (Women): 252
- No. of Principal Components: 300
- Explained Variance: 97.8%
- Stratified 10-Fold Cross
- Validation Accuracy: 59.2%

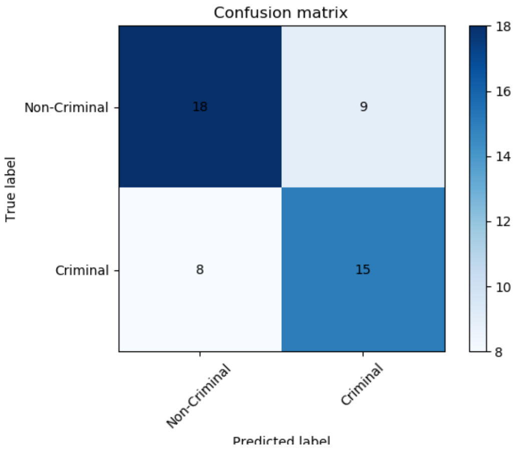

Figure 8: Confusion Matrix - Criminal Men/Non-criminal Women

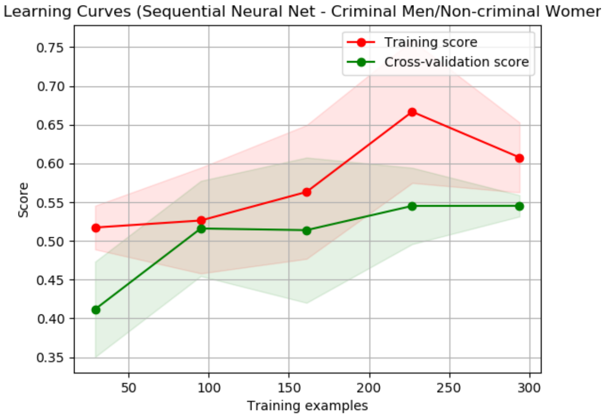

Figure 9: Learning Curve Criminal Men/Non-criminal Women

5.3 Results Set 3 Criminal Men vs. Non-criminal women

(Mixed race and emotion)

- Criminal images (Women): 77
- Non-criminal images (Men): 78
- No. of Principal Components: 120
- Explained Variance: 98%
- Stratified 10-Fold Cross Validation
- Accuracy: 51%

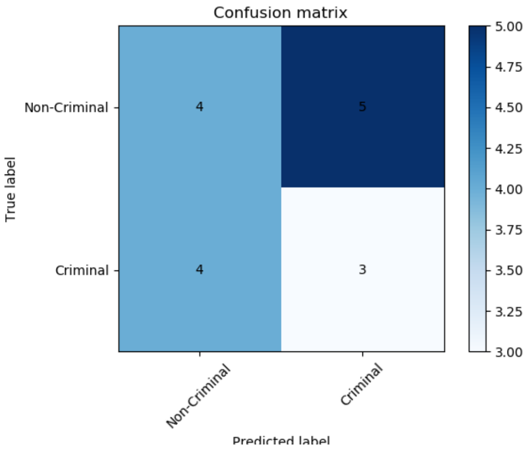

Figure 10: Confusion Matrix - Criminal Women/Non-criminal Men

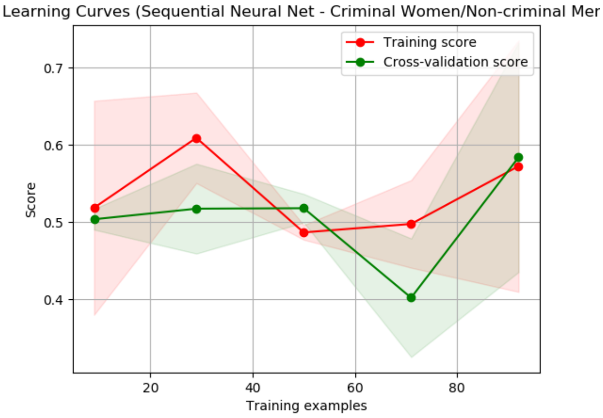

Figure 11: Learning Curve - Criminal Women/Non-criminal Men

. . .

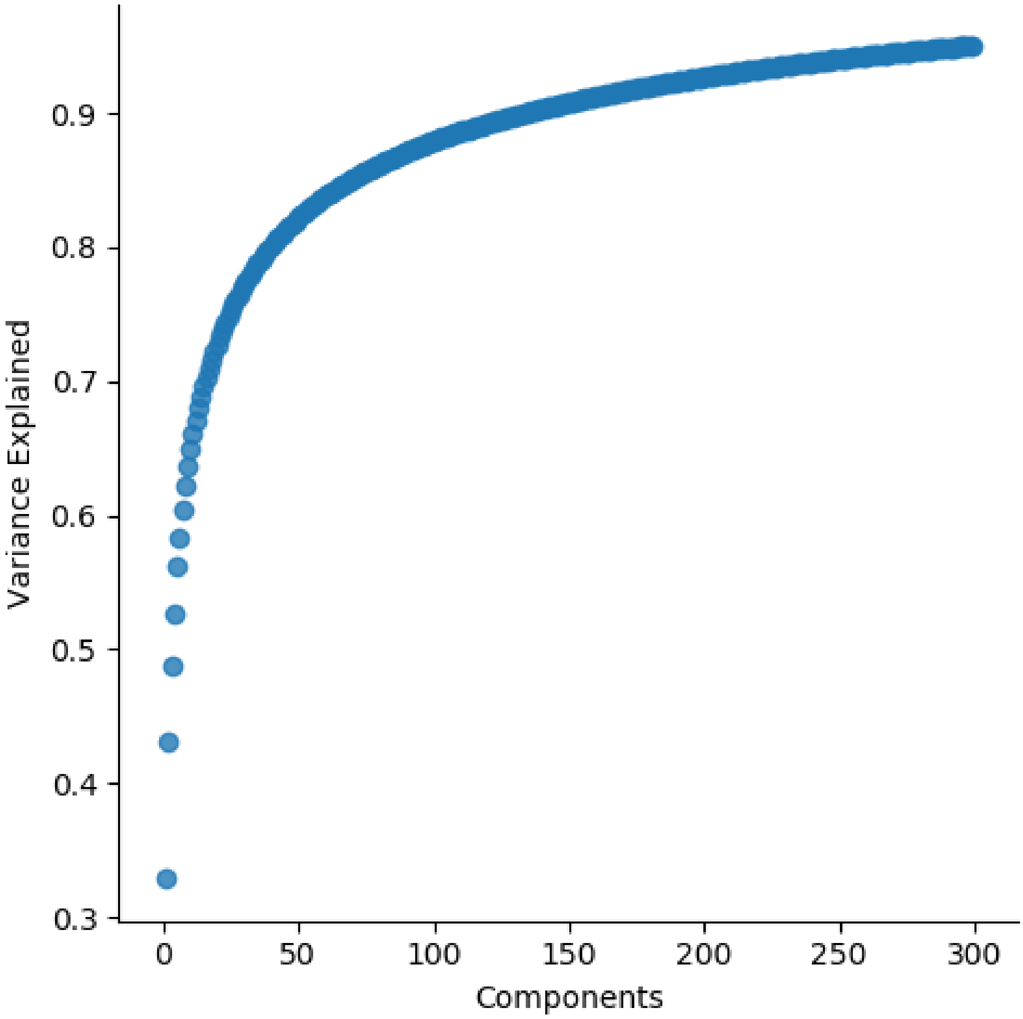

Figure 12: Principal Components Graph

5.4 Results Set 4 Principal Components Graph

figure 12. above

...

5.5 Results Set 5 Emotion Classification of Image Sets

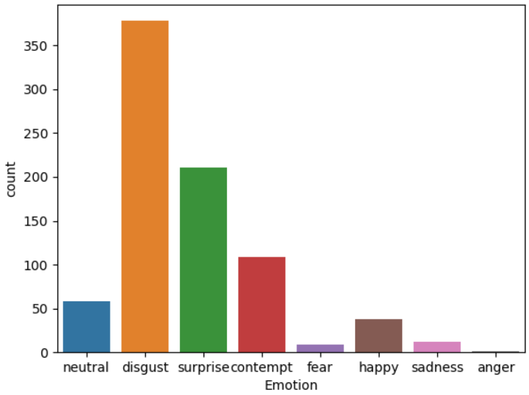

Figure 13: Emotion Profile (Criminals)

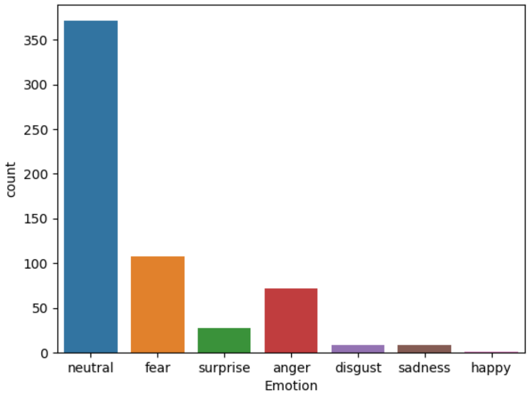

Figure 14: Emotion Profile (Non-criminals)

...

5.6 Discussion

The first scenario that was investigated was the full image sets i.e. 481 non-criminals and 621 criminals. The 10-fold stratified cross validation accuracy for this was 60.8%. For all scenarios, the confusion matrices were calculated on a train-test split of 90:10. The confusion matrix for this scenario shows an overall accuracy of approximately 75% predicting 21 of the 68 criminals as non-criminals and 7 of the 43 non-criminals falsely as criminals. The learning curves generally exhibit increasing accuracy as no. of training examples increase. The cross-validation learning curve accuracy increases from 0.5 to 0.58 over nearly 700 training examples. Precision is 36/(36+21) = 63.1% while recall is 36/(36+7) = 83.7% The next scenario is the criminal men (240) vs. non-criminal women (252). The 10-fold cross-validated accuracy is 59.2%, nearly on a par with the accuracy on the full image sets. However, the number of images in this case is 492 in comparison to 1102, for the full image set. This reduction in training data may be significant for the learning model and the fact that the accuracy is very close to the full set accuracy suggests that the accuracy of the 492 image set could exceed that of the 1102 set (60.8%) if it had a comparable number of images. The 492 image set may be performing nearly as well as the 1102 set due to the images in each set being exclusively from different genders. This is a potential bias in any classifier which is training on sets with differing gender balance. The classifier may be discriminating on the different structure of male and female faces as opposed to some latent criminality. The last tranche of images investigated was the case of female criminals and male non-criminals. Unfortunately, the number of images which satisfied this criterion was 77 criminal women and 78 non-criminal men. The 10-fold cross-validated accuracy was found to be 51%. In this case, the accuracy is difficult to compare to the other scenarios in terms of biases. This is primarily due to the image sets being small. 155 images (77 + 78) is not much to train a classifier on and consequently we would expect poorer performance. It was hoped to be able to perform emotion classification on the both the criminal and non-criminal dataset with a view to balancing the emotions in each set, thus excluding emotion bias. From histographic analysis of the image sets we can see in Figure 13 that the emotion detector model classified a large proportion of the criminal images as negative (disgust and contempt). Surprise was also prevalent. In opposition to this, the model classifies the non-criminals as dominantly neutral with some anger and fear. The result of this is that there is little overlap in the emotion profiles of the two image sets and it was not possible to create a scenario where image sets were emotion balanced before being presented to the criminal classifier model. The procurement of more images would further support the emotional bias investigation.

6 Conclusion and Future Work

We must be careful when attempting to classify people in any manner but the concept of predicting whether a person is a criminal must surely warrant the closest of scrutiny and rigorous analysis. The cost of misclassification may be very great indeed. The models that were designed and built within this research project support the investigation of biases across image sets. The Deep Learning neural net model for classification of criminals suggested that higher accuracy may be achieved if the gender balance is skewed across data-sets. The emotion classifier model suggests that the emotion profile across data-sets can be vastly different, most likely depending on the context of the situation. Such an imbalance in facial emotions may well have impacted any accuracy of a model searching for latent criminal tendencies. Suggested future work includes the procurement of a larger image set which would greatly support robust validation and bias investigation. It would also be potentially useful to utilize a convolutional neural network in place of the Sequential model while the emotion detector model would also benefit from more training data. The use of other dimension reduction techniques would also be interesting to investigate e.g. kernel PCA or a supervised technique such as Linear Discriminant

Analysis.

References

A., K. and Sutskever, I. (n.d.). Imagenet classification with deep convolutional neural networks, **1**.

Al-Rfou, R., Alain, G. et al. (2016). Theano: A Python framework for fast computation of mathematical expressions, *arXiv e-prints* **abs/1605.02688**.
URL: *http://arxiv.org/abs/1605.02688*

Ekman, P., . F. W. V. (1978). The facial action coding system (facs), a technique for the measurement of facial action, *Consulting Psychologists Press* .

K. Dabov, A. Foi, V. K. and Egiazarian, K. (2007). Image denoising by sparse 3-d transform-domain collaborative filtering application to statistical process control cascade of simple features., *IEEE Transactions on Image Processing* **8**: 20802095.

Kutne, R., Konugurthi, P., Agarwal, A., Rao, C. R. and Buyya, R. (2001). Computer vison, *Softw., Pract. Exper.* **46**(1): 79–105.

M.-H. Yang, D. J. K. and Ahuja., N. (2002). Detecting faces in images:a survey, *IEEE Transactions on Image Processing* **1**: 34–58.

Oliver, N., P. A. . B. (n.d.). A real-time face and tracker with facial expression recognition. pattern recognition, *Softw., Pract. Exper.* **1**(1).

Reece, A. G. and Danforth, C. M. (2016). Instagram photos reveal predictive markers of depression, *CoRR* **abs/1608.03282**.
URL: *http://arxiv.org/abs/1608.03282*

Saporta, G. and Niang, N. (1996). Principal component analysis: application to statistical process control cascade of simple features., p. 123.

Tsapatsoulis, N., K. K. S. G. P. F. . K. S. (2000). A fuzzy system for emotion classification based on the mpeg-4 facial definition parameter set.

Turk, M. A. and Pentland, A. P. (n.d.). Face recognition using eigenface.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features.

Wang, W., Tang, B., Fan, X., Mao, H., Yang, H. and Zhu, M. (2017). Efficient visibility analysis for massive observers, *Procedia Comput. Sci.* **111**(C): 120–128.
URL: *https://doi.org/10.1016/j.procs.2017.06.018*

Wang, Y. and Kosinski, M. (2017). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.
URL: *osf.io/zn79k*

Wu, X. and Zhang, X. (2016). Automated inference on criminality using face images.

X.Geng and Yin, C. (2013). Facial age estimation by learning from label distribution, *IEEE., Trans. Pattern Anal. Mach.* **35**(1): 79–105.

Zebrowitz, L. A. (1998). Social psychological face perception: Why appearance matters, *Social and Personality Psychology Compass* **2**: 1497–1517.

Zhang, Y., . Q. J. (2005). Active and dynamic information fusion for facial expression understanding from image sequences, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**(1): 27,699–714.

Zheng, X. Y. (2005). Person face examination based on adaboost algorithm cascade of simple features., pp. 167–169.