NIPS 2018 Competition proposal: Bias in Face-based Attributes Analysis as a function of skin type

Esube Bekele* Joy Buolamwini Timnit Gebru
Wallace Lawson
esube.bekele.ctr@nrl.navy.mil

February 14, 2018

0.1 Overview of the competition

Despite recent success in face recognition and facial attributes analysis, research suggests the existence of a wide range of unintentional biases towards a specific racial or gender group [Phillips *et al.*, 2011] [Klare *et al.*, 2012]. Facial analysis is increasingly becoming pervasive in our daily lives for various applications such as face-based verification, identification, and criminal suspect profiling. These biases have significant ramifications as these automated facial recognition and facial attributes analysis systems are widely adopted by law enforcement and other critical areas such as healthcare. Most common facial attributes in automated facial analysis include gender, race, and age [Fu *et al.*, 2014] [Ng *et al.*, 2015] [Han *et al.*, 2015]. Face-based Gender and facial skin tone classification are covered in this competition.

A recent study on effect of demographics on facial gender classification showed this bias on three commercially available gender classification systems [Buolamwini and Gebru, 2018]. As part of eliciting these biases, a new dataset was collected that is balanced in gender and skin tone type called the Pilot Parliaments Benchmark (PPB). Although there are many publicly available face datasets that have race, gender, age,and other face-based attributes, we will primarily use this benchmark dataset for evaluation (as test set) in this competition as attempts were made to balance the dataset based on demographic groups. This pilot analysis showed that three commercial face-based gender recognition systems by Microsoft, Face++, and IBM resulted in error rates of 23.8%, 36.0%, and 31.1%, respectively, in dark skin type female samples from the PPB dataset. In contrast, the three systems had an average error rate of 0.53% for light skinned color males despite the fact that only 30.3% of the benchmark were light skinned males[Buolamwini and Gebru, 2018].

*The Leader organizer should be the first author of the proposal.

The competition will have two separate tracks. In the first track, researchers will be generating samples from existing datasets to illustrate such racial and gender biases in existing facial attributes analysis commercial systems and methods. In the second track, we will provide imbalanced dataset for race and gender and researchers will submit solutions that learn the facial attribute while minimizing the gender and racial bias in the classification performance. We chose gender classification for two reasons. First, the original PPB dataset and pilot audit of the commercial systems is done on gender recognition and it will be easier for comparisons of the results coming out of this competition with the pilot study. Secondly, gender recognition involves recognition of several facial attributes such as facial shape, hair style and certain facial features and hence gender recognition implicitly is comprehensive.

Contributions: The two tracks of this competition has the potential to spur new research lines or new application of existing methodology in alleviating unintended algorithmic bias due to imbalanced datasets specifically for face-based gender and other attributes recognition.

- The first track is aimed at auditing algorithmic bias in commercial facial analysis systems and pinpointing the source of the bias (specifically if the bias is solely due to skin tone). Skin tone manipulation is a specific instance of image morphing in the color space. This could potenially attract solutions involving both traditional segmentation and manipulation and recent generative (and/or adversarial) for skin tone manipulation.

- The second track is aimed at developing solutions that would learn facial attributes from natural imbalanced and unconstrained datasets. As facial and full-body attributes recognition becomes increasingly important for soft-biometrics, watch-list surveillance [Kamgar-Parsi *et al.*, 2011], and person of interest identification and re-identification [Best-Rowden *et al.*, 2014], predictions that are equitable towards all demographic groups are certainly crucial. We hope to see several algorithmic solutions including, but not limited to, sample loss weighting, architectural innovations for skewed datasets, probability calibrations, few-shot or single-shot recognition, and learning the tail distribution.

0.2 Keywords

Other-race Effect, Automated Face Analysis, Gender Recognition, Bias in Facial attribute Analysis

0.3 Novelty

To the best of our knowledge this will be the first time such competition that is targeted at eliciting racial and gender biases in face recognition and analysis systems and

attempting to solve these biases using techniques other than simply balancing the training dataset. Balancing datasets for every demographic group and gender is usually impractical for commercial applications with consequential impact such as watch-list surveillance [Kamgar-Parsi *et al.*, 2011] and person of interest identification [Best-Rowden *et al.*, 2014].

1 Competition description

1.1 Background and impact

In psychology, there is a well documented phenomenon called "own-race" or "other-race" bias that people recognize faces and face related attributes of their own race better than that of other races [Furl *et al.*, 2002]. This inherent bias manifests itself in automated face recognition and analysis systems in part due to the unintentional imbalance in the datasets used to train such systems [Phillips *et al.*, 2011]. As shown in the Face Recognition Vendor Test (FRVT) by the National Institute of Standards and Technologies (NIST) evaluated every four years, the performance of automated facial analysis has been steadily improving [Grother *et al.*, 2010]. The first breakdown of facial analysis systems performance by demographic groups [Phillips *et al.*, 2011] showed, however, that this performance improvement is not uniform across demographic groups of race and gender.

With the wide spread adoption of automated facial analysis systems by law enforcement such as state and local police and federal immigration enforcement, proper regulation of such systems is of paramount importance. A recent report by the Georgetown law school center on privacy and technology highlights the presence and consequences of racial bias embedded in such systems [Garvie, 2016]. The report indicated that 50% (and expanding fast) of American adults are in a law enforcement face recognition network. It also highlights the dire consequences of such unregulated law enforcement face recognition showing such systems disproportionately affect African Americans as these systems are less accurate in identifying African Americans. Recent studies show the effect demographic groups on face analysis performance [Han *et al.*, 2015] [Farinella and Dugelay, 2012] [Klare *et al.*, 2012]. Although these preliminary findings seem to suggest the existence of such bias and its potential consequences, there is dearth of research in this area to both find the source of such bias in facial analysis systems and solutions.

It is important to pinpoint whether such bias is primarily due to skin tone (race or ethnicity) and gender or these systems are exploiting other background information such as clothing and other visible attributes that correlate with specific demographic groups. It is also important to solve this bias in facial analysis systems in the presence of an imbalanced training dataset. Therefore, this competition has two tracks correspondingly.

The first track requires participants to use the PPB dataset to generate samples that elucidate the source of such bias in commercial face recognition software. To generate these samples, competitors will synthetically manipulate the demographic dimensions of gender and facial skin tone of faces in the PBP dataset. The original PPB dataset is

balanced with respect to these demographics groups to avoid additional bias in the testing set [Buolamwini and Gebru, 2018].

In the second track, we solicit submissions that focus on a combination of novel and existing algorithms and architecture that result in relatively uniform performance across demographic groups. For this track, we will add additional samples as an extended test set to be used in combination with the PPB dataset. Esube has experience on data science competitions (with a master rank on kaggle.com) and we intend to use kaggle.com for hosting the competition. Joy and Timnit has collected and annotated the original PPB dataset.

1.2 Data

Demographic imbalances in existing face-based gender recognition datasets led to creating a new dataset for gender classification as evaluation benchmark called the Pilot Parliaments Benchmark (PPB) [Buolamwini and Gebru, 2018]. This dataset will be primarily used as an evaluation set for the preliminary rounds of the competition. The dataset is composed of 1270 identities (1 image/subject) from 3 African (Rwanda, Senegal, and South Africa) and 3 European countries'(Iceland, Finland, Sweden) national parliaments These countries were chosen by their ranking on percentage of women representation in their national parliaments. Fig. 1.2 shows representative samples and average images from each country and gender. Table 1 summarizes the PPB dataset statistics compared to recent face recognition benchmarks.

The major defining characteristics of facial analysis and recognition benchmark datasets is their imbalance with respect to racial/ethnic and gender demographic groups [Phillips *et al.*, 2011, Han *et al.*, 2015]. In the PPB dataset, intentional effort was made to balance among these demographic groups to allow a balanced evaluation of facial analysis algorithms (see Table 2). The countries included in the dataset were selected because they were among the top 10 countries based on their percentage of women representation in their national parliaments. This balance in the test dataset (PPB) sets equivalence across the demographic groups in all the metrics evaluated.

Table 1: Dataset statistics of PPB compared to IJB-A and Adience face recognition benchmarks

feature	PPB	IJB-A	Adience
Release Year	2017	2015	2014
# of Subjects	1270	500	2284
Avg. IPD (in pixels)	63	-	-
BBox Size	141 (avg)	≥ 36	-
Image Width	160-590	-	816
Image Height	213-886	-	816

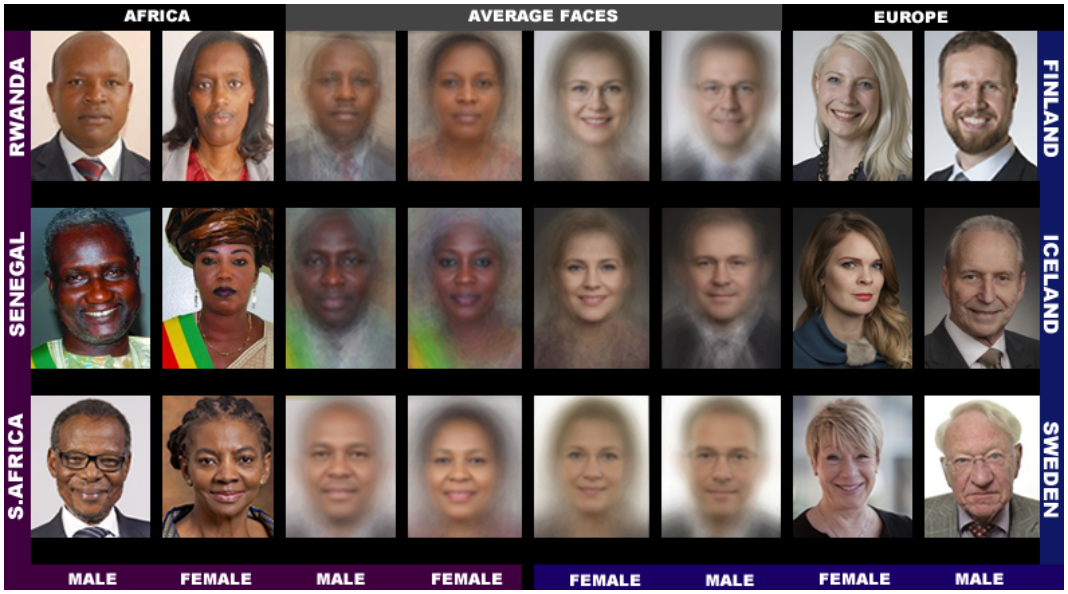

Figure 1: Sample images and average faces from each country and gender in the PPB dataset.

For the second track of the competition, we will use the CelebA dataset [Liu *et al.*, 2015] as a training set and the PPB dataset for preliminary evaluation on gender classification. While CelebA is a large unconstrained images of faces in the wild, PPB is a more balanced and constrained, in terms of pose and illumination, evaluation benchmark. CelebA contains 10,000 identities with 20 images each on average for a total of 202,599 images labeled with 40 attributes and 5 key points. The attributes include gender and skin color (only for pale skin) and hair color. Hence, it represents a typical imbalanced and unconstrained training set for facial analysis.

For the first track, we will release half of the African and half of the European set of images for a total of 634 images as part of the development package to generate morphed images by manipulating the skin tone and gender dimensions. We reserve the remaining

Table 2: Facial skin ton and gender breakdown of the dataset PPB compared to IJB-A and Adience

Demographics Group	PPB	IJB-A	Adience
Darker Female	21.3%	4.4%	7.4%
Darker Male	25.0%	16.0%	6.4%
Lighter Female	23.3%	20.2%	44.6%
Lighter Male	30.3%	59.4%	41.6%

636 images for the second track again released as part of the development package. We are extending this dataset by collecting and annotating 2500 additional parliamentarians' images from countries that are not already included in the original PPB and we will sequester this set as the final round test set for both tracks of the competition.

1.3 Tasks and application scenarios

This competition will have two separate tasks that are divided into two separate tracks. Competition participants are free to submit to both tracks.

1.3.1 Commercial Face Analysis Systems Challenge

In this challenge, participants are required to evaluate the performance of three commercial face-based gender classification systems (Microsoft, IBM, and Face++) with respect to the skin tone continuum. The main purpose of this challenge is to show the effect of skin type on gender classification. This is analogous to generating adversarial examples but constrained only by changing only the skin type. This challenge helps to pinpoint whether demographics (specifically skin type in this case) alone influence performance of gender classifiers and to rule out other confounding factors such as visible portions of clothing that would correlate with these specific demographic groups.

In this track, each participant will be given 634 images from the PPB dataset as part of a development kit that also contains tools for cross validation, pre-trained gender classifiers, links to the commercial face analysis cloud-based tools and evaluation performance metrics. For each image in the development package, each participant is expected to generate new image that is similar to the original image in identity. The only modification participants are allowed to make to the input images is to manipulate the skin tone of the subject in the image using any face morphing or feature-level facial attribute manipulation models.

The final performance in this challenge will be evaluated with a separate test set of images from parliamentarians from other countries to minimize over-fitting. In this task participants are prohibited to perform any adversarial manipulation on the image to get lower score other than the skin tone manipulation. Submission that results in the lowest performance in metrics as described in the metrics section will be winners.

1.3.2 Face-based Gender Recognition Challenge

This second task is aimed at training gender recognition models with an unconstrained non-balanced realistic dataset producing a balanced performance in each of the four course demographic categories. For this purpose, each participant will be required to train a model (or ensemble of models) on CelebA dataset. Participants are allowed to use all the 39 attribute labels as long as it helps improve the gender classification. For instance, they could use the "Pale_skin" attribute to categorize the images based on skin lightness and use sample weights during training. Participants are allowed to employ any architecture,

training technique, data augmentation (no external dataset is allowed), output probability calibration and other methods for training on imbalanced dataset.

We will release 636 images of the PPB dataset with the development kit for local cross validation and testing. We will collect and annotate new set of final test set images of parliamentarians from other countries than the original six countries in the original PPB dataset to avoid over-fitting.

In both tracks of the competition, participants are required to submit their code as part of their submission by the deadline date to be considered for the first three top position.

1.4 Metrics

We will use the metrics mean accuracy (see Eqn. 1) and F1 score (see Eqn. 4) as evaluation metrics for this competition mainly. Mean accuracy is more informative than regular accuracy (or error) metric on the balance of the accuracy in both values of the gender class (i.e. male and female). F1 score balances the precision (see Eqn. 2) and recall (see Eqn. 3) and hence could serve as a more balanced performance metric than just either precision or recall.

$$mA = \frac{|TP|}{|P|} + \frac{|TN|}{|N|} \tag{1}$$

where TP and TN are the true positive and true negatives over total positive samples and total negative samples of the gender attribute.

$$Prec = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|f(x_i|} \tag{2}$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i|} \tag{3}$$

$$F1\ score = \frac{2.Prec.Recall}{Prec + Recall} \tag{4}$$

where N is the total number of samples considered at one time and $|.|$ is the set cardinality.

For the first vendor evaluation track of the competition, we will re-score the updated vendors' systems in terms of F1 score (see Eqn. 4) using the original PPB dataset for the four demographic groups (dark skin female, dark skin male, light skin female, and light skin male) and will average the F1 scores as baseline. Images will be generated by running the source code of each submission on the evaluation set. The evaluation set will be used as testing benchmark to score the three vendor systems (Microsoft, IBM and Face++). This will be evaluated on each of the four demographic groups using both mean accuracy and F1 score, the lowest average F1 score across the three vendors and the four demographic

groups will determine the winners. If there is close to tie on the average F1 score, mean accuracy will be used to separate the teams.

For the second balanced gender recognition challenge, we will generate 4 F1 scores for the four demographic groups for each submission on the sequestered extended PPB benchmark. We will then sort submissions based on average F1 scores across the four groups. We will cut off the top 10 submissions (models) that have the higher average F1 scores. Although F1 score is a more balanced performance metric, averaging across the four demographic groups could dilute the 'other-race' effect. Hence we will perform Freidman's ANOVA test on the difference between the probability submitted for each sample in each group and the ground truth, i.e. $|y_i - f(x_i)|$ for the ith sample. The test will produce the probability that there is a difference across the four demographic groups, $p - value$ and χ^2 statistics that shows the consistency of prediction errors across the four groups. We will use this χ^2 statistics to re-sort the top 10 teams for consistency and this will determine the winners for the second part of the challenge.

1.5 Baselines and code available

We will release a self-contained development kit that contains, the necessary datasets needed for each track (see Section 2.1), the implementation of the evaluation metrics described in the Metrics section (see Section 1.4), and the baseline results.

For the commercial vendors audit section, the baseline will be the preliminary results presented in [Buolamwini and Gebru, 2018]. After this preliminary study two of the evaluated vendors (Microsoft and IBM) updated their gender recognition systems for better performance in the dark skin female category. We will evaluate the new systems and update the baselines when we release the development kit.

For the balanced gender classification task, our baseline model is ResNet50 pre-trained on ImageNet and fine-tuned on CelebA for gender recognition as part of the development kit. We will train the model, evaluate it on PPB based on the split explained in the Protocol section (see Section 2.1) and release the results as the baseline benchmark within the development kit. We will also release a pytorch-based code for training such a model on the CelebA dataset for gender recognition.

1.6 Tutorial and documentation

Participants will be encouraged to read [Buolamwini and Gebru, 2018] and the paper will be included in the development package as a starting point to understand the problem. Then, we will release extensive documentation and tutorial examples for how to evaluate the vendor benchmarks once participants have generated the skin tone modified images using the method of their choice. we will also thoroughly document the gender recognition pytorch code using CelebA. We will add a tutorial example showing how to evaluate the trained model on the local preliminary round part of the PPB evaluation benchmark. The

final test set of the extended PPB dataset that is currently being collected and annotated will NOT be released until the competition is over.

To summarize, the development package will contain the following:

- The pilot study paper wrote by two of the authors (Joy and Timnit) of this competition proposal that explains the problem [Buolamwini and Gebru, 2018].
- Two splits of the PPB dataset as evaluation benchmarks for the preliminary rounds of the two tracks of the competition.
- Tutorial and documentation on how to perform the vendor evaluation once participants have generated an altered image for the first track of the competition.
- Pytorch code to fine-tune an ImageNet trained ResNet50 (participants would change this to their architectures and their techniques to combat data inter-class imbalance), the trained model weight, and through documentation
- Example tutorial that shows how to evaluate their trained documents on the split set of the PPB dataset for their local cross-validation.

2 Organizational aspects

2.1 Protocol

Each track of the competition will have preliminary rounds of up to 5 submissions maximum and a final round submission. The preliminary rounds are for the purpose of giving participants feedback other than their local internal validation using the preliminary test sets released together with the development kit. The final round will be scored on the private test set and only this final round of submissions are considered for ranking teams.

We plan to apply for hosting research based competition with kaggle.com. It has self contained on-line submission and leader boards.

2.1.1 Preliminary Rounds

For the commercial face-based gender recognition systems evaluation track, we will release 634 random images from the PPB split by countries for internal cross validation and preliminary rounds submissions. The protocol for the preliminary rounds of this track are as follows.

1. Participants are expected to generate a new sample for each of the 634 images in the development kit and submit the generated image. Participants are allowed to use their own image manipulation models or traditional pipeline. No commercial software is allowed.

2. Organizers will score the newly generated images on the three commercial face-based gender recognition systems and performance metrics for each vendor and average performance metrics will be released for each of the participants. At this stage, organizers will not check for cheating.

For the face-based gender recognition challenge, participants will follow the following protocols.

1. Organizers will release 636 random images from the PPB evaluation benchmark as part of the development kit.

2. Participants are expected to train their face-based gender recognition models on CelebA dataset. The aligned and cropped versions of the CelebA dataset together with 40 attributes (gender is one of these attributes) will be made available both on the competition site and kaggle.com. We encourage submissions that exploit attributes other than gender to help in balanced gender recognition in a multi-label classification. Only the gender prediction will be used to score submissions.

3. Organizers will score the submissions on the 636 preliminary round test sets and release the gender recognition performance to participants (in the form of public leader board).

2.1.2 Final Rounds

For the final round we will set aside 2500 images and their skin type and gender labels for scoring the final submissions. To prevent cheating in this final round, the following protocols are followed:

- The labels for the final round images are held confidentially and will not be made available until the competition ends and winners are announced. The images will be released

- All participants will have to submit their code or open source it on github and submit a link to it to be considered for the top 3 positions. Submissions without source code in the final round will not be scored.

- For the first track of this competition, organizers will make at most care to run the code and generate new samples for the 2500 final test images. Special care will be taken for the top 3 teams to make sure that their code is not generating any adversarial samples other than skin tone manipulation in an effort to exploit adversarial vulnerability of these commercial gender recognition systems. After the skin tone manipulation, the original image and the generated image will not be L_∞ consistent and hence it is difficult to check if there is added adversarial attack on the generated images.

- For the gender recognition challenge track, organizers will run the classification code submitted by participants on the 2500 images and score them against their labels.

In this final round, the protocol for each track is as follows.

1. For both tracks of the competition, participants are required to submit their code by the competition deadline.
2. Organizers will run the source code on the 2500 final round test images and generate manipulated images and evaluate the three commercial systems with them in the case of the first track. For the second track, organizers will run the code and generate gender predictions and score them against the ground truth gender labels.
3. Organizers will determine the top 3 teams in each track.
4. For the first track, the top 3 submissions are selected based on performance metrics for each vendor and average lowest performance metric for each of the two demographic groups (i. e., dark skin type male and dark skin type female).
5. The source code from the top three teams will be carefully examined to make sure adversarial attacks are not injected together with the skin tone manipulation.
6. For the second track, the top 3 teams are selected based on the highest balanced performance metrics for all the four demographic groups.
7. Organizers will announce the results and release the final round 2500 test images and their labels. Organizers will combine the new 2500 test images with the original PPB dataset so that further research and commercial systems will use this as evaluation benchmark test sets for gender and race balanced facial analysis.

2.2 Rules

1. Anyone can participate with the exception of the organizers and anyone that has any conflict of interest with the organizers.
2. To be eligible for scoring in the final rounds, participants are required to submit their source code or open source it on github.
3. Source code must not produce any errors or the submission will be abandoned without scoring.
4. Participants are required to submit clear documentation of their code on how to run it on the final round test images. The documentation also should clearly list set of dependencies for the code.

5. Any leader board probing will not be allowed and will be grounds for elimination from the competition.

6. External datasets are prohibited. Only the CelebA dataset and PPB dataset released by the organizers are allowed for this competition.

7. Participants are allowed to pre-training with ImageNet only. No other datasets may be used.

2.3 Schedule

The proposed competition schedule is as follows.

April 1st, 2018 **Start of competition promotion**: Start of Competition Promotion. We will launch the competition website with the call for submissions. Start heavy promotion of the competition via social media, at upcoming conferences and mailing lists.

June 30th, 2018 **Release of development kit and official start of the competition**: The competition starts with the release of the development kit.

June 30th, 2018 - October 1st, 2018 **Duration of the competition**: Submissions are allowed starting this date until the competition closes for a maximum of 5 preliminary round of submission by each team for each track.

October 1st, 2018 **Deadline for the final round**: Each team will have one more final round submission which is due on October 1st, 2018.

October 1st 2018 - October 30th, 2018 **Final round evaluation**: We will evaluate final round submissions on the separate test images that are not publicly released.

November 1st 2018 **Announcement of winners**: Winners (top 3 teams) in each track will be announced and the final round test images will be made public.

2.4 Competition promotion

This competition will be promoted using the organizers' Facebook, Twitter, Reddit and other social media accounts and pages, and mailing lists. Moreover, we encourage black professionals in AI to take part in this competition. To this effect, we will heavily promote it in the Black in AI Facebook group and discussion forums and we wish winners of both tracks to present in the next Black in AI workshop which will be co-located with NIPS 2018.

2.5 Organizing team

Esube Bekele: is a National Research Council Fellow at the US Naval Research Lab (NRL) in the Navy Center for Applied Research in Artificial Intelligence (NCARAI). Esube has extensive experience in data science kind of competitions and he currently serves in the organizing team od DC Data Science (Large meetup for data science professions in the

Washington, DC area). Currently, he holds a master rank at kaggle.com. Esube will serve as the primary coordinator of the competition providing data, development kit, platform administration, baseline methods and evaluation performance metrics–all with in the scope of his regular work as fellow at NRL.

Joy Buolamwini:

Timnit Gebru:

Wallace Lawson: is a Research Scientist at the US Naval Ressearch Lab (NRL) in the Navy Center for Applied Research in Artificial Intelligence (NCARAI). He has research and published extensively on facial analysis, including image morphing, open-set face recognition and attributes prediction. Wallace will serve as an evaluator and will review submissions to the competition. As mentor to Esube at NRL, Wallace also will oversea the competition progress and will serve as an advisor.

3 Resources

3.1 Existing resources, including prizes

We hope to host the competition at kaggle.com similar to last year's NIPS competition on adversarial examples. We do not plan to provide monetary prizes. However, top 3 submissions in each track will be invited to present in Black in AI workshop, which will be collocated with NIPS 2018 and this competition track of NIPS 2018.

References

[Best-Rowden *et al.*, 2014] Lacey Best-Rowden, Hu Han, Charles Otto, Brendan F Klare, and Anil K Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security*, 9(12):2144–2157, 2014.

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[Farinella and Dugelay, 2012] Giovanna Farinella and Jean-Luc Dugelay. Demographic classification: Do gender and ethnicity affect each other? In *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*, pages 383–390. IEEE, 2012.

[Fu *et al.*, 2014] Siyao Fu, Haibo He, and Zeng-Guang Hou. Learning race from face: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2483–2509, 2014.

[Furl *et al.*, 2002] Nicholas Furl, P Jonathon Phillips, and Alice J OToole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26(6):797–815, 2002.

[Garvie, 2016] Clare Garvie. *The Perpetual Line-Up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

[Grother *et al.*, 2010] Patrick J Grother, George W Quinn, and P Jonathon Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST interagency report*, 7709:106, 2010.

[Han *et al.*, 2015] Hu Han, Charles Otto, Xiaoming Liu, and Anil K Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1148–1161, 2015.

[Kamgar-Parsi *et al.*, 2011] Behrooz Kamgar-Parsi, Wallace Lawson, and Behzad Kamgar-Parsi. Toward development of a face recognition system for watchlist surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1925–1937, 2011.

[Klare *et al.*, 2012] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[Ng *et al.*, 2015] Choon-Boon Ng, Yong-Haur Tay, and Bok-Min Goi. A review of facial gender recognition. *Pattern Analysis and Applications*, 18(4):739–755, 2015.

[Phillips *et al.*, 2011] P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):14, 2011.