Towards causal benchmarking of bias in face analysis algorithms

G. Balakrishnan∗,‡, Y. Xiong‡, W. Xia‡, P. Perona†,‡
∗ Massachusetts Institute of Technology
†California Institute of Technology
‡ Amazon Web Services

Supplementary material to ECCV2020 paper
Extended version available on arXiv: `https://arxiv.org/abs/2007.06570`

A Face Attribute Annotation in Synthetic Images

The face images used in our experiments are synthetic, and therefore there is no real person behind each image. Thus, there is no intrinsic ground truth for face attributes such as gender, hair length, and skin tone. Such attributes are instead established by human annotators. We clarify here what we mean when we talk about face attributes in the absence of a physical ground truth.

Many attributes have both intrinsic and extrinsic manifestations. For example, "emotion" may be studied at three levels [3]: an unconscious physiological state, conscious self-perception (feelings), and emotional display (e.g. facial expression) [12]. These quantities are *intrinsic* to a person's or an animal's body and are not directly accessible to a machine. By contrast, an *extrinsic* description, i.e., the report by an onlooker of his/her perception, are more easily accessible, and this is what the machine is trained to predict.

Since we are using synthetic images, it should be clear that we are not attempting to access the intrinsic state of a person: there is no person, and there is no intrinsic gender, ethnicity, age or emotion. However, perception of such attributes is possible. This is the same way that onlookers instinctively classify the *Venus of Milo* as "female" and Michelangelo's *David* as "male," despite the fact that they are idealized marble representations, rather than real people.

Thus, when we refer to the "age" or "gender" or any other attribute that is computed by a face analysis system from a picture, what we mean is *the algorithm's prediction of a casual observer's report of their perception of the outwards display of that attribute*. This is a bit of a mouthful, and that's why we use the abbreviated expression of "attribute," "age" or "gender." The attributes we measure from human observers are reports of subjective perceptions. However, as we find in Sec. 3.3, these measurements are consistent and reproducible across different observers, and so we consider statistics of such reports as objective quantities.

In our study, we discretize continuous face attributes. We have used six classes of age and skin tone, five of hair length, facial expression and gender, etc. (see Figs. 13 and 15). This choice was made to conform with the literature, e.g., the Fitzpatrick scale of skin tone [17], and to accommodate the abilities of non-expert casual observers, the "common person," whose perception we rely on in our experiments. We make no claim to have the perfect discretization scheme; other discretization choices may be better suited in different contexts.

Gender deserves a special mention: *gender identity* is often modeled as multi-dimensional [15]. However, here we are measuring *reports of gender perception* (an extrinsic variable), rather than gender identity (the intrinsic variable), and our subjects could not reliably report beyond the traditional one-dimensional M/F dimension. Therefore, following [10] we settled for one dimension, which we discretized into five steps to accommodate different levels of confidence and ambiguity.

B Method

Algorithm 1: K-attribute transect generation

Input: Generator G, tuples $\{(L_k, \mathbf{n}_k, b_k, \mathbf{v}_k, \mathbf{c}_k)\}_{k=1}^{K}$, where L_k is a transect dimension, $(\mathbf{n}_k, b_k)$ is a hyperplane, $\mathbf{v}_k$ is a direction vector, and $\mathbf{c}_k$ are signed decision values.
Output: A $L_1 \times \cdots \times L_K$ transect T^i.

$\mathbf{z}^i \sim p(\mathbf{z})$
$\mathbf{z}^{i,0}$ = projection of $\mathbf{z}^i$ onto intersection of $\{(\mathbf{n}_k, b_k)\}_{k=1}^{K}$
for $l_1 = 1 \cdots L_1$ **do**
  $\vdots$
  for $l_K = 1 \cdots L_K$ **do**
    $T^i(l_1, \cdots, l_K) = G(\mathbf{z}^{i,0} + \sum_{k=1}^{K} \frac{\mathbf{c}_k[l_k]}{\langle \mathbf{v}_k, \mathbf{n}_k \rangle} \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|})$

Algorithm 2: Orthogonalization

Input: Vectors $\{\mathbf{n}_j\}_{j=1}^{N_a}$.
Output: Vectors $\{\tilde{\mathbf{n}}_j\}_{j=1}^{N_a}$, where $\tilde{\mathbf{n}}_j \perp \mathbf{n}_k, \forall k \neq j$

$Q, R \leftarrow$ QR-factorization of matrix $[\mathbf{n}_1, \mathbf{n}_2, \cdots, \mathbf{n}_{N_a}]$
for $i = 1 \cdots N_a$ **do**
  $\tilde{\mathbf{n}}_i = \mathbf{n}_i$
  for $j = 1 \cdots N_a$ **do**
    if $i \neq j$ **then**
      $\tilde{\mathbf{n}}_i = \tilde{\mathbf{n}}_i - \frac{Q_j \cdot \langle Q_j, \tilde{\mathbf{n}}_i \rangle}{\langle Q_j, Q_j \rangle}$

Fig. 12. 1D transects with and without orthogonalization. Without orthogonalization (see Sec. 3.1), decreasing hair length results in more masculine-looking faces. This phenomenon is not as apparent after orthogonalization (Sec. 3.1). We see only slight orthogonalization differences in the skin color transects, indicating that the skin color hyperplane was already near-orthogonal to the other attribute hyperplanes.

C Annotations

Each annotator evaluated each image for one attribute at a time. For each image, we collected 5 annotations per attribute for a total of 40 annotations per image. We discretized each attribute using three to six levels.

The number of annotations that are needed by our method is rather formidable. However, we found that this is not an obstacle in practice. In our experiments, $\mathcal{D}_{\mathbf{z}}$ consists of 5,000 images, and $\mathcal{D}_{transect}$ consists of 1,000 8-image transects (see examples in Fig 5). The total number of annotations was thus 13,000 (images) x 8 (attributes) x 5 (annotations per image and per attribute) = 0.52M annotations. Amazon Mechanical Turk delivered on average 10-20 annotations per second, thus annotations took about 10 hours to complete over two separate sessions. Annotators were paid 1.2c per annotation, earning 10-15 US$ per hour.

Fig. 14 shows the raw annotations for one 1D transect and three attributes. One may see that there are very few outlier annotations, and that in most cases annotations fall in one or two neighboring attribute levels.

Fig. 15 (top left) shows a distribution of per-image annotation standard deviations, split by attribute. One unit corresponds to the dynamic range of each attribute. For most attributes, the median annotator standard deviation is near 0.1, i.e. less than the separation between attribute levels. These observations indicate good agreement between annotators and suggest that annotations are meaningful and reproducible.

Fig. 15 (top right) presents the distribution of mean annotator fakeness scores for the synthesized images. Only a small portion of images are deemed "Likely fake" or "Fake for sure." Realism of images is particularly important in our analysis, since image artifacts can unknowingly affect the decisions of gender classifiers. In our experiments, we remove images with a fakeness score above a certain threshold (see Sec. D.1).

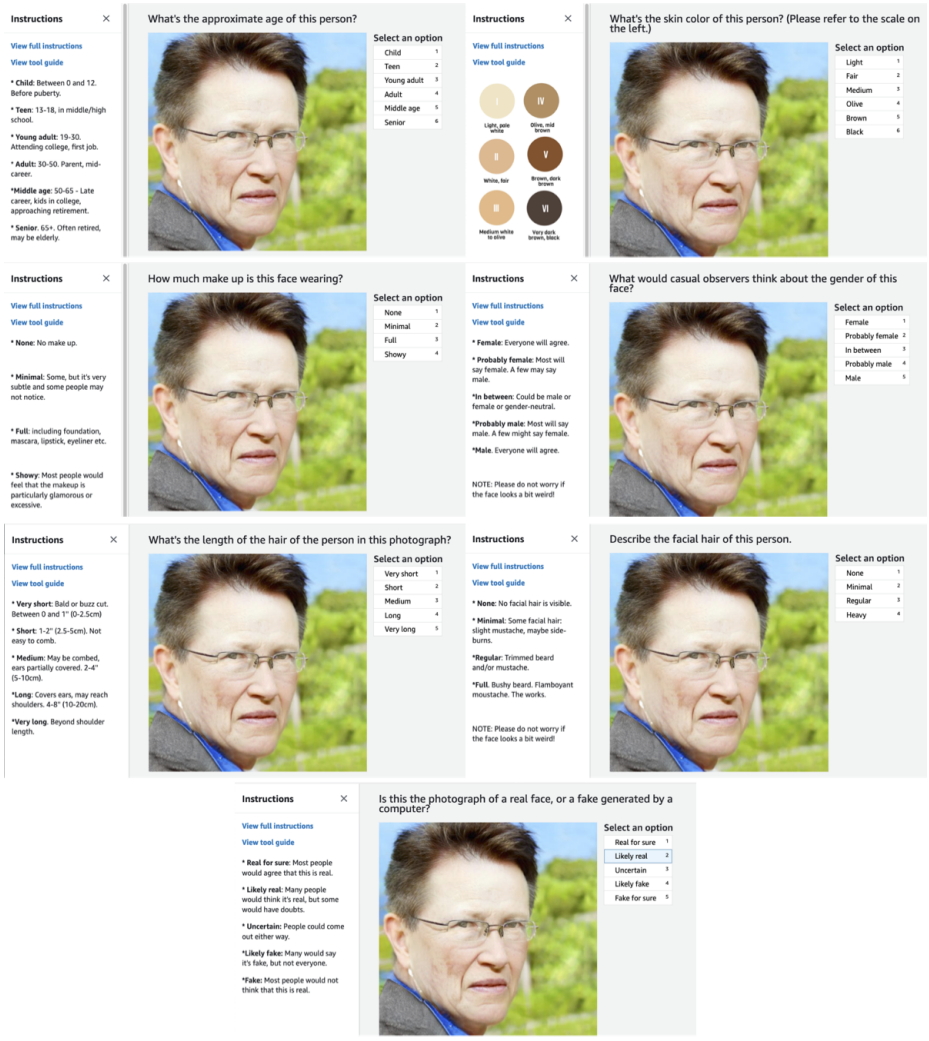

Fig. 13. Screenshots of the graphical user interface for seven annotations we collected from Amazon Mechanical Turk annotators using the SageMaker Ground Truth service [1].

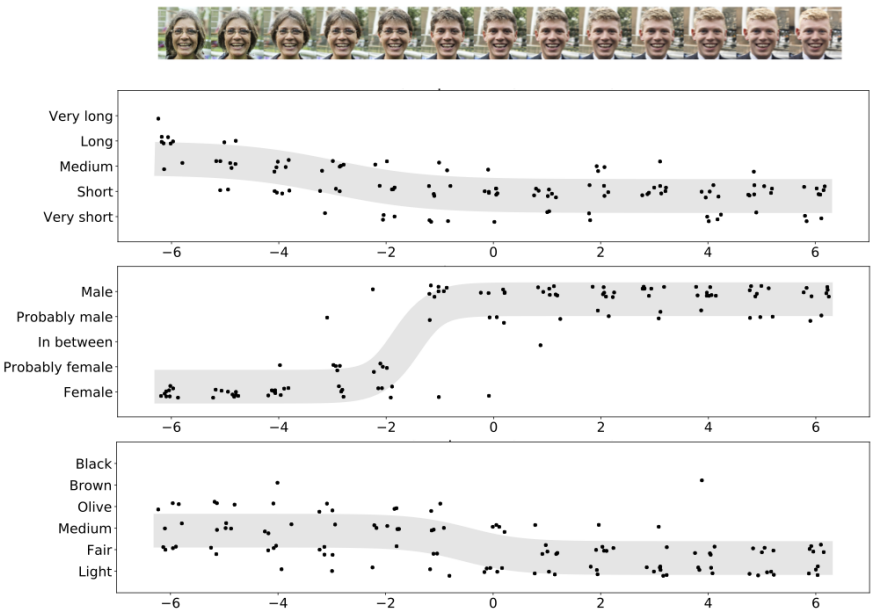

Fig. 14. Annotation consistency. Hair length (top), gender (middle) and skin tone (bottom) annotations on a 13-image 1D transect. This transect was annotated in a pilot experiment to fine tune our GUIs and to evaluate the consistency of the annotators, and not used in our main experiment. Here nine annotations were obtained for each attribute and for each image. The annotations are shown as dots below each image. The x axis increments one unit from one image to the next. A small amount of noise was added in x and y in order to visualize the individual annotations. The thick gray curves show the fit of a logistic function to the data. Annotations typically fall within one or two neighboring attribute levels. There are very few outliers. For a quantitative overall analysis see Fig.15.

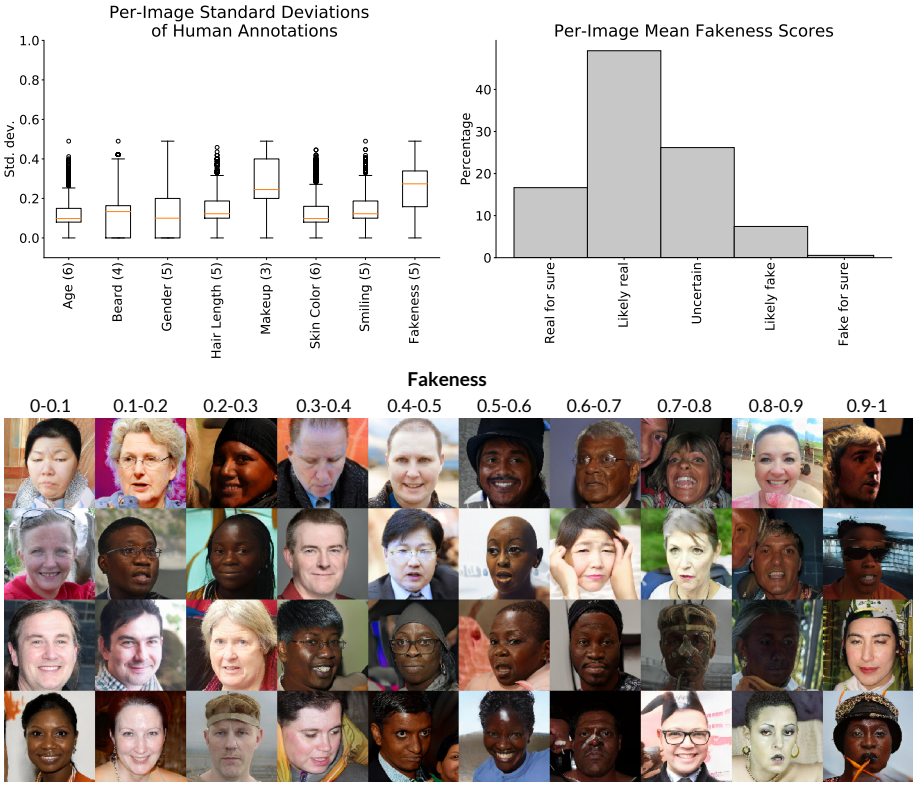

Fig. 15. Annotation quality and image realism. (Left) Distributions of per-image standard deviations of human annotations for each of the attributes we considered (one unit = dynamic range of the attribute). Five annotators were asked to provide a rating for each attribute of each image. The number of rating options per attribute is indicated in brackets next to the attribute's name. The median standard deviations (red lines) are comparable to the quantization step, indicating good annotator agreement. (Right) We asked our annotators to rate the realism of the images. The distribution of such scores is shown. Fewer than 10% of the ratings indicated fake or likely fake, suggesting that the synthetic images we randomly sampled are fairly realistic. (Bottom) we show examples of synthesized faces organized by mean human fakeness scores. Images with high fakeness scores were removed from the experiments (see Sec.D.1).

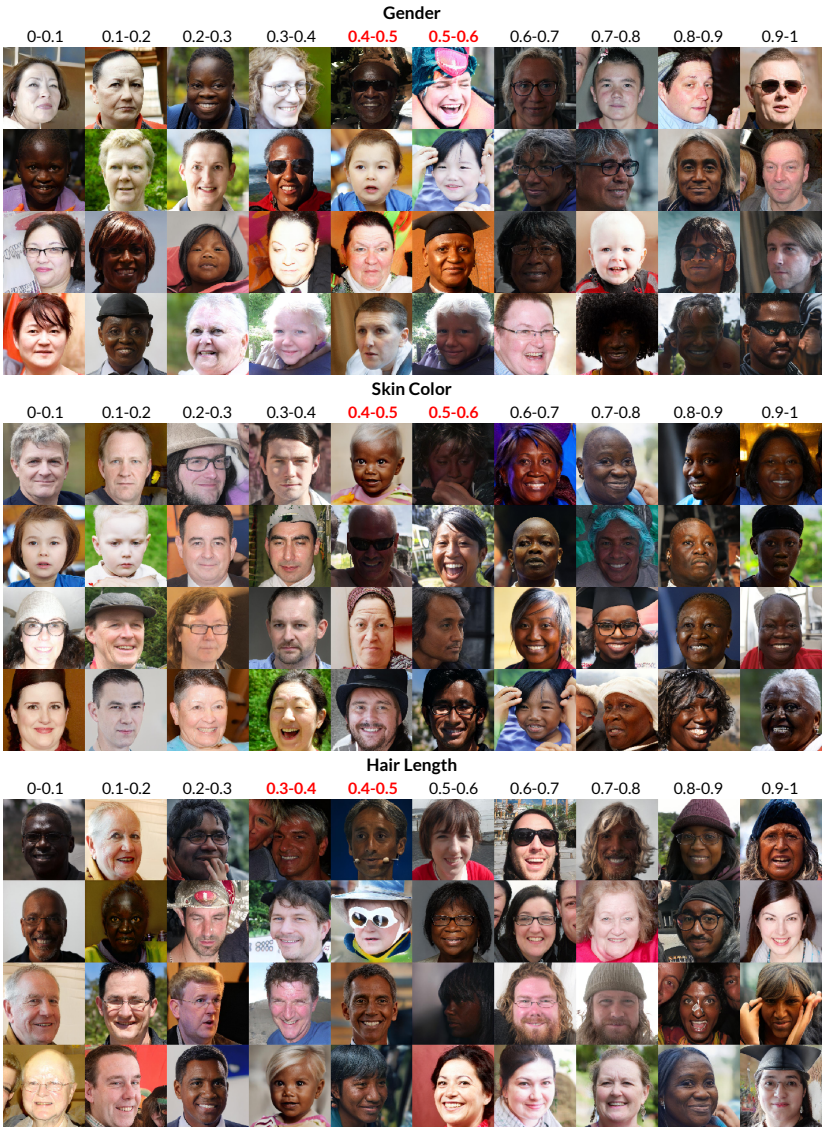

Fig. 16. Samples of synthesized faces, organized by mean human annotation scores. In our analysis, we omitted faces from ranges indicated in red to focus on clearly perceived females/males, light/dark skin tones, and short/long hair lengths.

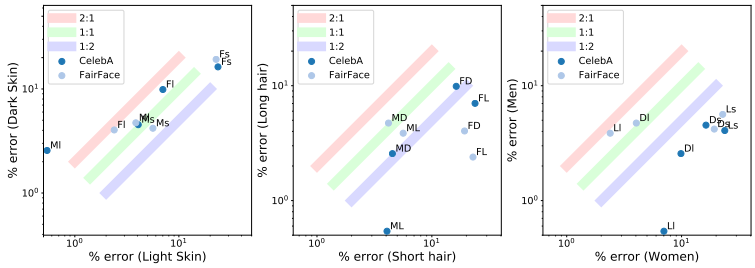

Fig. 17. Scatter plots of error rates using data from Fig. 8 (transects). Each dot compares the error rates of a pair of groups that differ by one attribute only (indicated in the label of the x and y axes). The two letters near each dot indicate the shared attributes ('M/F' indicate male and female, 'D/L' indicate dark and light skin, and 's/l' indicate short and long hair). Dots falling along the equal error line indicate that skin tone has little or no effect on error. In contrast, females and persons with short hair have higher error rates.

D Experiments and Results

D.1 Dataset Pruning

We remove any transect image with a mean fakeness score greater than or equal to "Likely fake" (0.75 in the normalized range of $[0, 1]$). We also removed faces with attribute values in the normalized subranges of $[0.4, 0.6]$ for skin color and gender, and $[0.3, 0.5]$ for hair length (see Fig. 16 for examples). After these pruning steps, we were left with 5713 images.

D.2 Gender Classifier Training

We trained our classifiers for 20 epochs with the binary cross-entropy loss. We set the learning rate at $1e^{-4}$ for the first 10 epochs, and $1e^{-5}$ for the final 10 epochs. To avoid a baseline bias of predicting one gender over another, we enforced the likelihood of sampling male and female faces during training to be equal.

D.3 Bias Analysis

Fig. 17 presents the same data in Fig. 8 for easier comparison of bias across intersectional groups.

D.4 Logistic Regression

We discretized attributes into levels, and assigned a binary variable to each level. We used the same discretization for hair length (short vs. long hair), skin color (light vs. dark skin) and gender (female vs. male) used in our experiments thus

far. We used two levels for beard (no/light beard vs. beard) and makeup (no/light makeup vs makeup), three for facial expression (serious/frown vs. neutral vs. smile), and the original semantic levels for age described in Fig. 13. In all, this resulted in 17 input variables to our logistic regression model. We used scikit-learn's LogisticRegression function [44], and set the regularization parameter to 1.

D.5 Joint Effects of Attributes

Fig. 17-right shows that error rates vary across different intersectional groups of skin color and hair length in a way that is not simply a linear combination of each attribute.

This is also the reason we removed children and teenagers from our analysis, as these individuals tend to have different appearance characteristics from adults. Fig. 19 illustrates this, by breaking down error rates by age and gender subpopulations for two classifier decision thresholds. The difference in error rates between the genders is fairly consistent for young adults to middle-aged individuals, but vary for children/teenagers and seniors. This demonstrates that age and gender have joint effects on errors.

Fig. 18 shows faces from our synthesized transects on which the ResNet models were most incorrect. For each gender misclassification direction, we show faces on which the model predictions were farthest from the average human annotator response. ResNet-CelebA tends to heavily misclassify young male children/babies as female, in line with the quantitative result in Fig. 19.

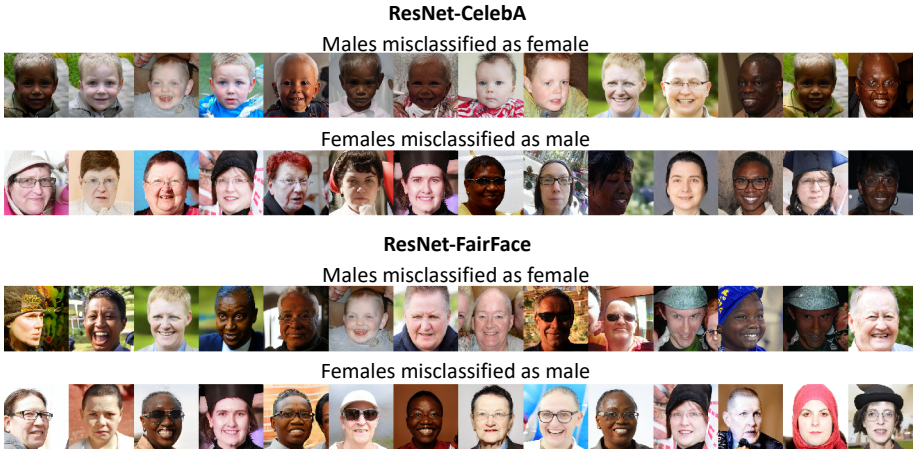

Fig. 18. Images with largest errors. Synthetic faces on which the classifiers most deviated from the mean human annotations.

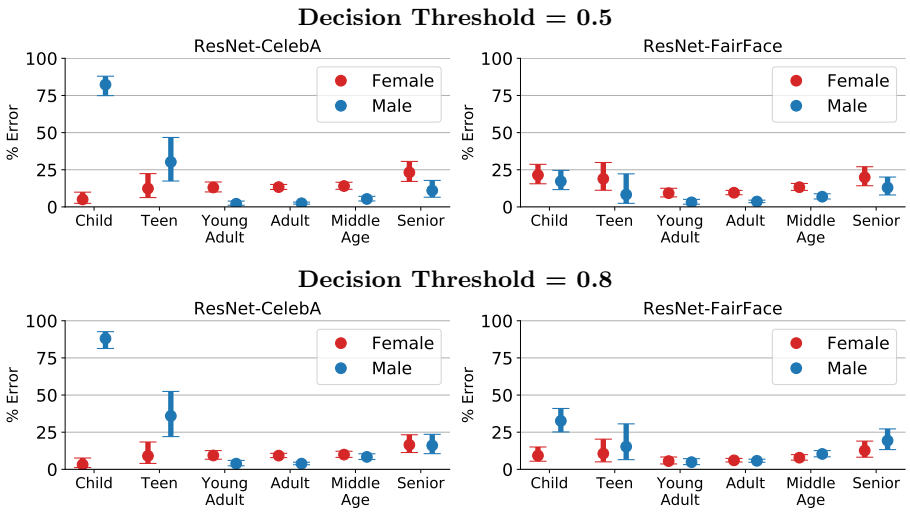

Fig. 19. Errors by gender and age group on our transect images. The two top plots were obtained by using a decision threshold equal to 0.5, and show a prevalence of female errors. The bottom two plots were obtained with a threshold equal to 0.8, chosen to minimize overall error. There is a non-uniform influence of age on errors. Both models tend to have lower errors for young to middle-aged adults. The differences in errors between genders are fairly consistent for adults, but differ for children, teenagers and seniors, illustrating a combined age-gender bias in the algorithms.