UNIVERSITY OF MIAMI

Beyond Black and White

Assessing Deep Learning Facial Classifications by considering Race and Ethnicity as a Multidimensional Physical Characteristic

Rahul Dass

Ph.D. Student, U-LINK Predoctoral Fellow

April 6, 2020

Dr. Ubbo Visser (UM)
Dr. Nick Petersen (UM)
Dr. Marisa Omori (UMSL)

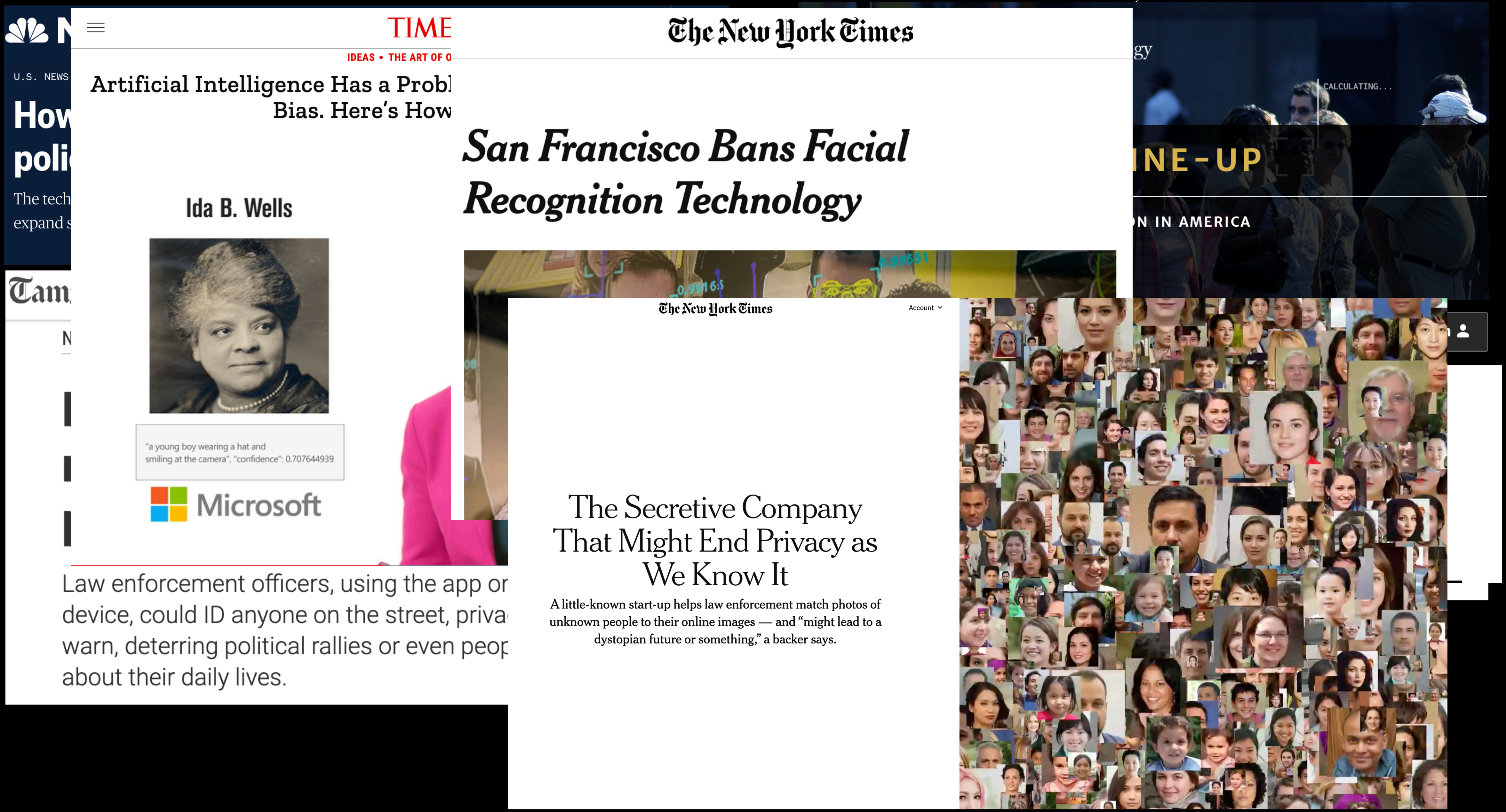
[Source: Algorithmic Justice League]

Facial Processing Technology (FPT)

Broadly encompasses various facial classification tasks:

- **Detection** of the face and facial landmarks (eyes, nose, etc.)
- **Analysis** of the face (age, gender, race/ethnicity, etc.)
- **Recognition** of the face (identify or verify)

FRT/FPTs’ Issues in Society



Rise of Fairness, Accountability and Transparency in ML

[Source: Time Magazine]

Outcomes / Inspiration / Consequences:

- Led companies to update their APIs (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019; Raji et al., 2020)
- Curating "less biased" benchmark datasets (Buolamwini and Gebru, 2018; Merler et al., 2019; Kärkkäinen and Joo, 2019)
- Investigate relationships between sensitive physical characteristics and demographic groups (Dwork et al., 2018; Ryu et al., 2018)

My Inspiration

- Given the lack of research concerning Hispanic face classification within computer vision, sociolegal and criminology communities…

- Across 13 CV papers, “Race” always seen to belong to *one* of several subcategories including White, Black, Hispanic, Indian, East Asian, Southeast Asian or Middle Eastern…

- From CRT, “Race” should not be considered simply as a singular defining attribute but as a *multidimensional* construct (Hanna et al., 2019)

Research Questions

- How would a DLM's performance vary if the classification task changed from race to race-ethnicity prediction using the same dataset?

- Does the performance of DLM race-ethnicity classifications vary based on the model architecture?

- Does the performance of these DLM tasks vary when using human annotations based on a single rater versus multiple raters?

Data and Interdisciplinary Methods (1/2)

- Analyzed a novel dataset of 194K MDC arrestees' mugshots (2010-2015)
- UM Sociology Student Raters Survey 14K stratified samples (29-labels) including:
  - Two Race (Black and White)
  - Four Race-Ethnicity (Black Hispanic, White Hispanic, Black Non-Hispanic, White Non-Hispanic)
  - Seven Skin Tone (type 1 or "very light" to type 7 or "very dark")
- Fill missing ethnicity labels in court data using "surnames text-based" approach (Word and Perkins, 1996; Wei et al., 2006; Word et al., 2008; Elliott et al., 2009; King and Johnson, 2016)

Table 1: Comparing U.S. and MDC General Demographic Spreads, 2010, vs. MDC Arrestees Population, 2010 – 2015

Race-Ethnic Subgroup	U.S. General	MDC General	MDC Arrestees
Black Hispanic	0.4%	1.9%	9.18%
White Hispanic	8.7%	58.4%	39.70%
Black non–Hispanic	12.2%	17.1%	37.96%
White non–Hispanic	63.7%	15.4%	13.14%
Total	**100.0%**	**100.0%**	**99.98%***

* Other racial–ethnic groups represented a very small (0.02%) proportion and were removed from the dataset.

[Source: Dass et al., 2020 – Forthcoming]

Data and Interdisciplinary Methods (2/2)

- Developed 7 DLMs using transfer learning based on ImageNet weights (fastai/PyTorch and Keras/TensorFlow)
- Varying experimental parameters:
  - Sample size (Balanced vs. Imbalanced)
  - Image Preprocessing (Raw vs. OpenFace)
  - Metric (Accuracy)
  - Hyperparameters (lr_finder)
  - Fine-tuning (freezing)

(a) Raw Black Mugshot (b) Raw White Mugshot (c) OpenFace Black Mugshot (d) OpenFace White Mugshot

[Source: Dass et al., 2020 – Forthcoming]

Results (1/3)

- Improved DLM prediction accuracies:
  - ✓ Race by 5.49%
  - ✓ Race-Ethnicity by 10.22%
- At a cost of annotating 100-times and 50-times more data – which would be an expensive process
- Given small number of skin tone samples, DLM performed poorly
- Co-presented at CCS Social Informatics Lecture Series called “Gigabytes for Good”

Table 2: DLM-based results for three classification tasks using ResNet-50

Sample Size	Classification Task		
	2 race	4 race–ethnicity	7 skin tone
Balanced*	91.72%	70.71%	63.97%
Imbalanced†	97.21%	80.93%	64.39%

* 1K samples per race and race–ethnicity subgroup; 399 samples per skin tone type

† Full dataset: 200K samples for race and race–ethnicity; Stratified dataset: 14K samples for skin tone

Results (2/3)

Table 2: Comparing the performance of 7 DLMs for binary (Black and White) race classifications based on court and student annotated mugshots, 2010-2015.

Model	**Raw Images**		**OpenFace**	
	Courts	Students	Courts	Students
ResNet–50	92.00%	93.50%	93.73%	91.72%
AlexNet	92.00%	92.75%	92.73%	89.72%
Inception–v4	94.25%	92.00%	93.98%	88.22%
SE–ResNet–50	93.75%	93.50%	93.98%	91.47%
SE–ResNext–50_32x4d	93.75%	89.25%	94.23%	89.72%
VGG–16_bn	94.00%	92.25%	92.23%	**93.98%**
VGG–19_bn	94.25%	92.50%	**94.48%**	91.47%

(a) Balanced classification: 1,000 samples per race subgroup.

Model	**Raw Images**	**OpenFace**
	Courts	Courts
ResNet–50	97.20%	**97.21%**
AlexNet	97.17%	96.84%
Inception–v4	97.26%	96.79%
SE–ResNet–50	97.37%	97.18%
SE–ResNext–50_32x4d	97.52%	97.12%
VGG–16_bn	97.45%	97.13%
VGG–19_bn	97.50%	97.08%

(b) Imbalanced classification: full Miami–Dade County arrestee population.

[Source: Dass et al., 2020 – Forthcoming]

- After 28-experiments, based on two label sources, DLMs achieved greatest accuracies of 94.48% (courts) and 93.98% (students) for a balanced dataset with OpenFace preprocessing
- No singular model architecture performed “the best” under all experimental settings
- Comparing VGG-19_bn (balanced courts) with ResNet-50 (imbalanced courts), find a gain of only 2.73% despite using approx. 100-times more data!

Results (3/3)

Table 3: Comparing the performance of 7 DLMs for four race-ethnicity classifications based on court and student annotated mugshots, 2010-2015.

Model	Raw Images		OpenFace	
	Courts	Students	Courts	Students
ResNet–50	56.20%	73.30%	55.31%	70.71%
AlexNet	58.75%	75.87%	60.95%	73.46%
Inception–v4	59.00%	71.25%	51.43%	67.83%
SE–ResNet–50	61.12%	76.25%	**61.32%**	**74.84%**
SE–ResNext–50_32x4d	61.25%	79.12%	48.31%	70.46%
VGG–16_bn	60.50%	76.37%	58.19%	74.09%
VGG–19_bn	63.87%	77.12%	59.57%	74.09%

(a) Four race–ethnicity classification: balanced (1,000) samples per race subgroup.

Model	Raw Images	OpenFace
	Courts	Courts
ResNet–50	80.60%	80.93%
AlexNet	79.09%	79.93%
Inception–v4	80.79%	80.18%
SE–ResNet–50	80.61%	**81.05%**
SE–ResNext–50_32x4d	80.40%	80.77%
VGG–16_bn	80.26%	77.92%
VGG–19_bn	80.43%	79.77%

(b) Four race–ethnicity classification: imbalanced full arrestee population.

[Source: Dass et al., 2020 – Forthcoming]

- Average OpenFace Court data across 7 DLMs, performed slightly better than chance (56.44%) – not helpful!
- Improved accuracies for imbalanced court DLMs is suspicious since 75% of data belonged to WH and BnH
- [Most Important] Student rated DLMs outperformed their court annotated counterparts consistently, ranging from 12.51% to 22.15% increase in accuracy.
- Balanced Student SE-ResNet-50 only underperformed by 6.21% than Imbalanced Court SE-ResNet-50

Model Inference – Validating

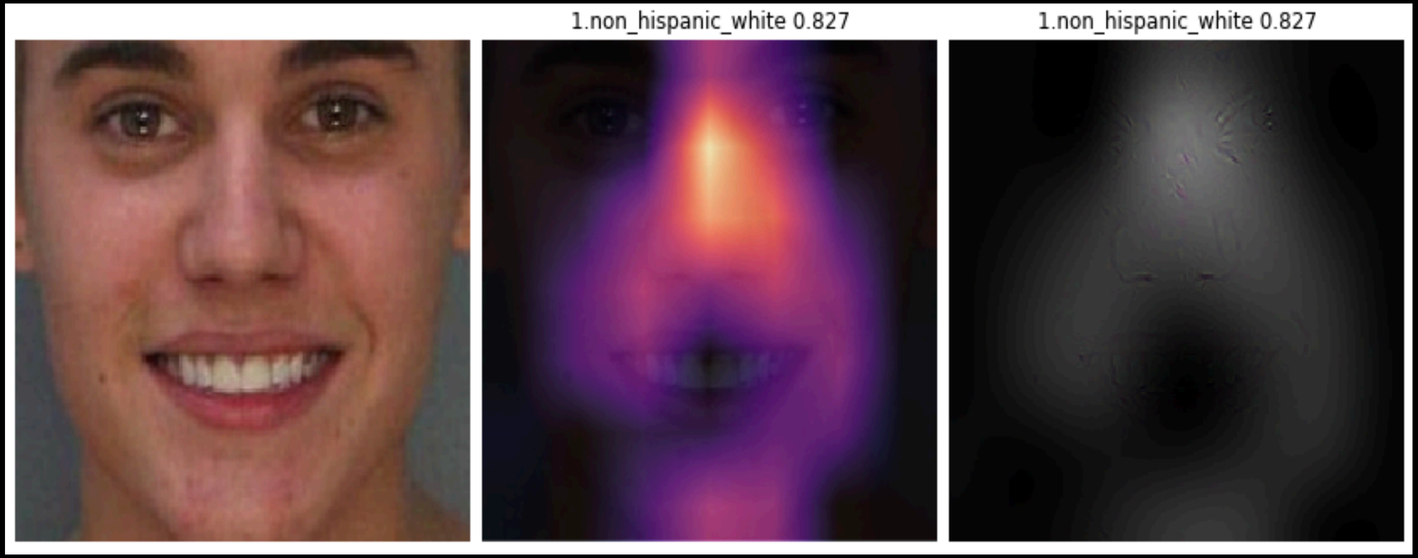
SE-ResNet-50 Model Inference – Testing

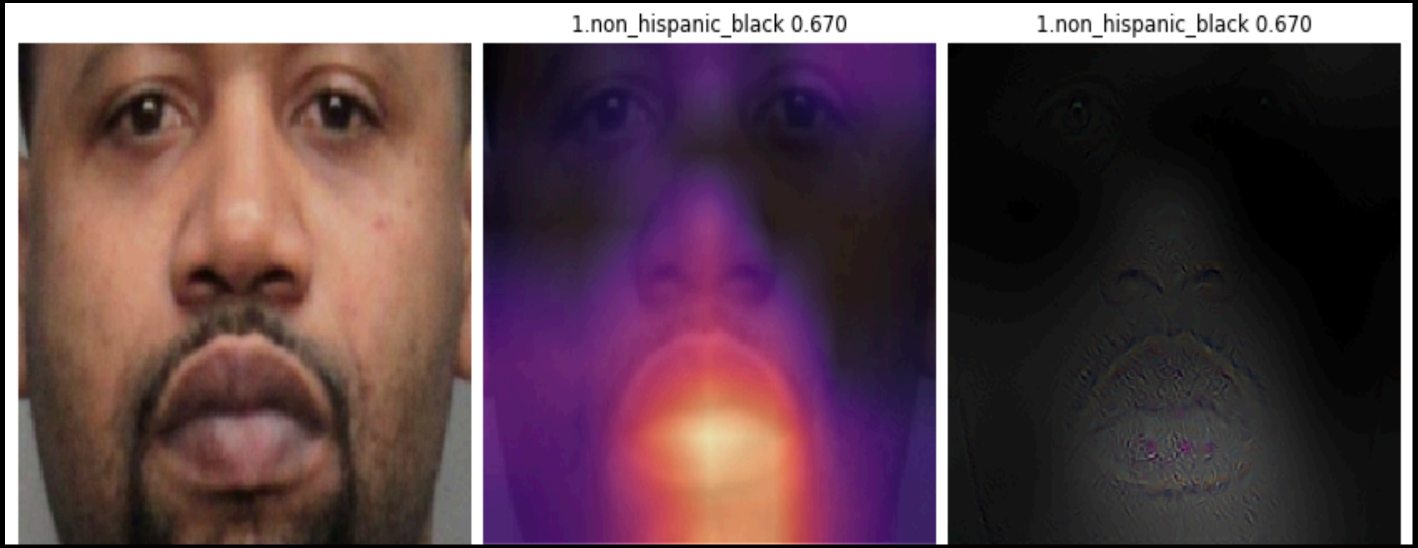
- Both mugshots were correctly classified:
  - Non-Hispanic White (82.7%)
  - Non-Hispanic Black (67.0%)

- Two heatmaps reveal:
  - Non-Hispanic White – structure centering about the nose
  - Non-Hispanic Black – structure centering around the (bottom) lips

- Despite being trained on a balanced race-ethnicity sample size, confidence for Black mugshot much lower than White counterpart

- Investigate if similar disparities exist for larger datasets





Future Work

- Given that ImageNet weights were used, investigate if training DLMs from scratch or models specifically with face weights makes a difference?

- Inference learning via “Balanced Student Race-Ethnicity” SE-ResNet-50 model:
  - Generate additional 190K DLM-based race-ethnicity labels and compare performance with Imbalanced “surnames text-based” Court trained SE-ResNet-50 (81.05%)

- Evaluate how biased each DLM is w.r.t. each race-ethnicity subgroup and assess if the new methodology fosters DLMs to be more demographically inclusive

Conclusions

- Novel multidimensional approach for understanding and annotating “race” in face datasets by looking at race-ethnicity combinations
- Achieved 74.84% accuracy for race-ethnicity using only 2% of the annotated dataset – “bigger is not always better”
  - Outperforming court records by 12.51% to 22.15%
  - Investigate implications in terms of court sentencing outcomes to suggest a new methodology for various interested communities
- Moving the literature forward particularly for Hispanics and working towards a more inclusive approach when building FPTs