It’s Not Just Black and White

Classifying Defendant Mugshots Based on the Multidimensionality of Race and Ethnicity

UNIVERSITY OF MIAMI

UMSL

Rahul Kumar Dass, Dr. Nick Petersen, Dr. Ubbo Visser
(University of Miami)

Dr. Marisa Omori
(University of Missouri-St. Louis)

AI-CRV 2020, Ottawa, Canada

May 12-15, 2020

Rise of Fairness, Accountability and Transparency in ML

[Source: Time Magazine, 2019]

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 (PPB, DiF, FairFace) (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)

Our 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 Critical Race Theory, “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)
- 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

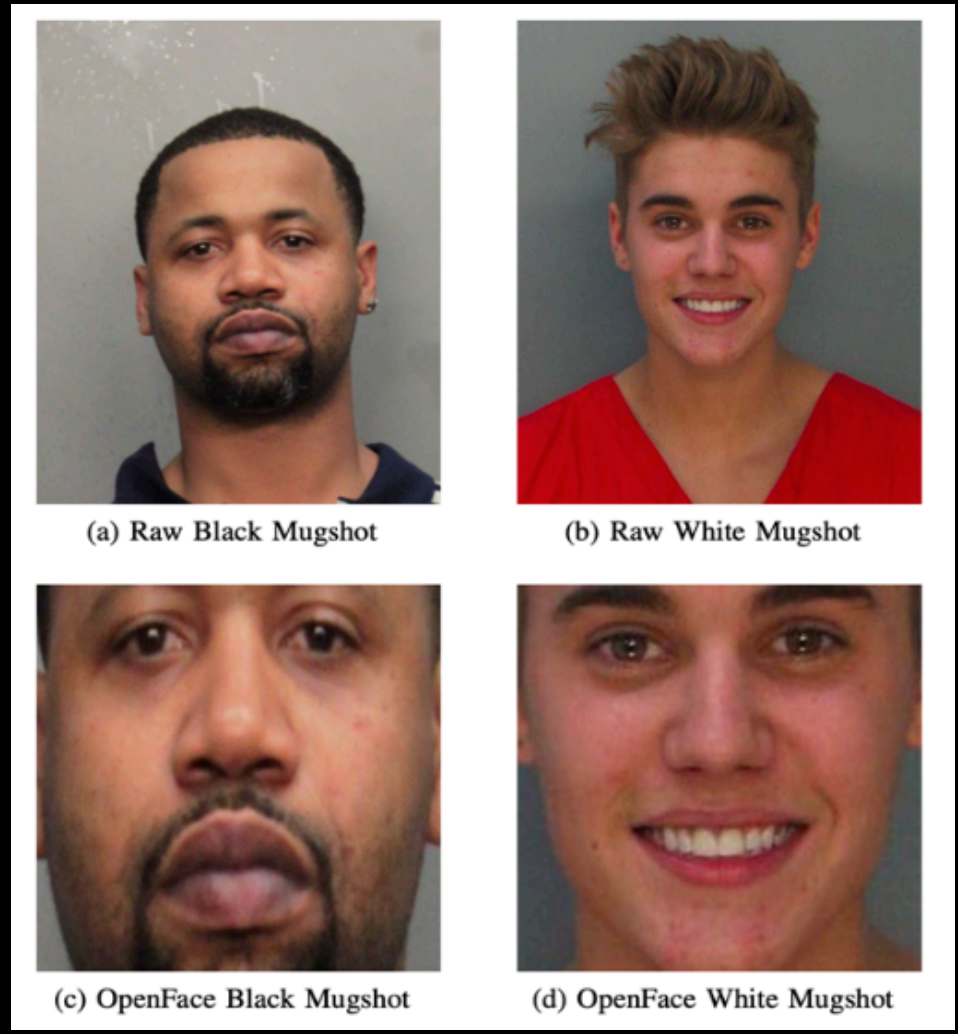
Race-Ethnic Subgroup	U.S. General	MDC General	MDC Arrestees
Black Hispanic	0.4%	1.9%	9.18%
White Hispanic	8.7%	58.4%	39.70%
Black non–Hispanic	12.2%	17.1%	37.96%
White non–Hispanic	63.7%	15.4%	13.14%
Total	**100.0%**	**100.0%**	**99.98%***

* 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)
- 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 (1/2)

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]

- DLMs achieved greatest accuracies of 94.48% (courts) and 93.98% (students) for a balanced dataset with OpenFace preprocessing - not much difference comparing single vs. multiple raters for binary classification
- No singular model architecture performed “the best” under all experimental settings
- Comparing VGG-19_bn (balanced courts ~ 2K) with ResNet-50 (imbalanced courts ~ 194K), find a gain of only 2.73% despite using approx. 100-times more data!

Results (2/2)

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 (~ 4K samples) only underperformed by 6.21% than Imbalanced Court SE-ResNet-50 (~ 194K samples)

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

Conclusion

- 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

Thank you!

[rdass@cs.miami.edu]