Research Track | Responsible Web

Perceptions in Pixels: Analyzing Perceived Gender and Skin Tone in Real-world Image Search Results

Jeffrey Gleason, Avijit Ghosh, Ronald E. Robertson, Christo Wilson

Northeastern University

1. MOTIVATION

Existing work: focus on hand-selected queries (e.g. “doctor”) to quantify racial and gender bias in image search [1,2]

Our work: analyzes **real-world image search queries**. What do people search for? How representative are results?

[1] Kay, M., Matuszek, C., & Munson, S. A. (2015, April). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 3819-3828).
[2] Metaxa, D., Gan, M. A., Goh, S., Hancock, J., & Landay, J. A. (2021). An image of society: Gender and racial representation and impact in image search results for occupations. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-23.

2. RESEARCH QUESTIONS

RQ1: What are the most popular categories of open-ended people queries**?

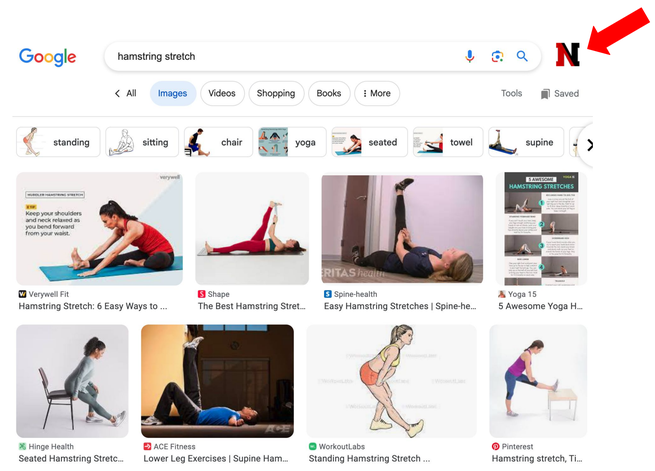
RQ2: How representative are results, with respect to. perceived gender, skin tone, and age?

RQ3: How does representativeness vary across search engines and categories?

RQ4: Do people refine their queries with demographic words (e.g. “women”)?

***Open-ended people queries:* queries that return images of people and are not predetermined (i.e. no named entities, no demographic adjectives)

3. DATA COLLECTION



We answer our questions by building a browser extension to collect over 50,000 unique image search queries on Google and Bing from over 600 US residents.

		Participants		
		N	%	US Census
Gender	Female	334	51.9	50.4
	Male	310	48.1	49.6
Race/Ethnicity	White	518	80.4	58.9
	Black	49	7.6	13.6
	Hispanic	34	5.3	19.1
	Asian	14	2.2	6.3
	Native American	1	0.2	1.3
	Two or more	13	2.0	3.0
	Other	15	2.3	–
Age	< 18	0	0.0	21.7
	18-64	507	78.7	50.4
	≥ 65	137	21.3	17.3

Table 2: Demographics of participants who contributed image search queries.

One important limitation is that these searchers skewed White and none were under 18.

			Images/Query	
Search Engine	**№ Screenshots**	**№ Images**	**Mean**	**Std**
Google Images	54211	2510331	46.31	8.09
Bing Images	54127	2688838	49.68	2.70

Table 3: Summary statistics from image search crawls.

We collected the top 50 image search results for each of the 50,000 queries on Google and Bing in summer 2022.

4. METHODS

	Filtering Step	№ Queries	Query Fraction	№ Users
	Original sample	54070	1.00	643
1.	>= 25% of images have people	21539	0.40	550
2.	Not named entity	4387	0.08	415
3.	Safe for work	3728	0.07	404
4.	Manual review	1481	0.03	296

Table 4: Summary of sample size after each filtering step.

We **identified open-ended people queries** by applying models for person detection, named entity recognition, and not-safe-for-work detection.

Category	Example Queries	№ Queries Sampled	Average № Faces/ Image
fashion	tuxedo, face mask, masks	20	1.2
military		20	1.6
politics		20	3.5
art	photo caption, tattoo	18	1.6
medicine	covid patients	18	1.5
sports	exercise, stretching	18	2.7
children		17	1.5
food	sitting	17	1.7
body care		15	1.9
psychology	conversation, optimistic	11	1.6
sexuality		11	1.9
telecomms		10	2.1
veterinary		10	1.4
pedagogy		8	3.1
tourism		7	3.0

Table 5: Example queries, number of sampled queries, and average number of faces per image in top 15 query categories.

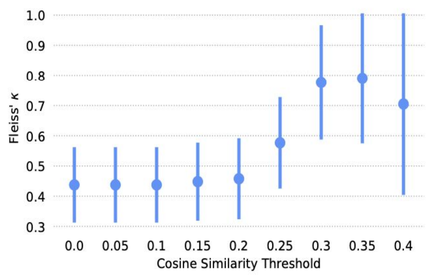


Figure 2: Category assignment agreement as cosine similarity threshold varies. Bars show 95% confidence intervals.

We **categorized open-ended people queries** into a WordNet taxonomy according to their cosine similarity with the category names.

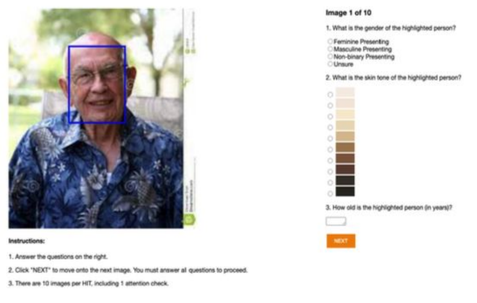


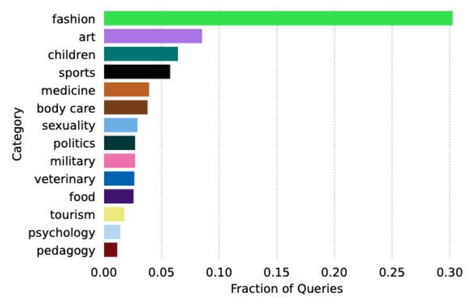
Figure 7: Mechanical Turk labeling interface.

Label Type	Fleiss' κ 95% CI	Weights
Gender Presentation	(0.81, 0.85)	Identity
Skin Tone	(0.44, 0.52)	Quadratic
Age	(0.79, 0.83)	Quadratic

Table 6: Fleiss' κ agreement statistics between labelers.

We sampled a subset of queries from each category from each category and designed a Mechanical Turk task to **label perceived gender, skin tone, and age**. We paid workers $14/hour.

5. RESULTS



RQ1: Fashion is by far the most popular category for open-ended people queries. Queries related to art, children, and sports are also relatively common.

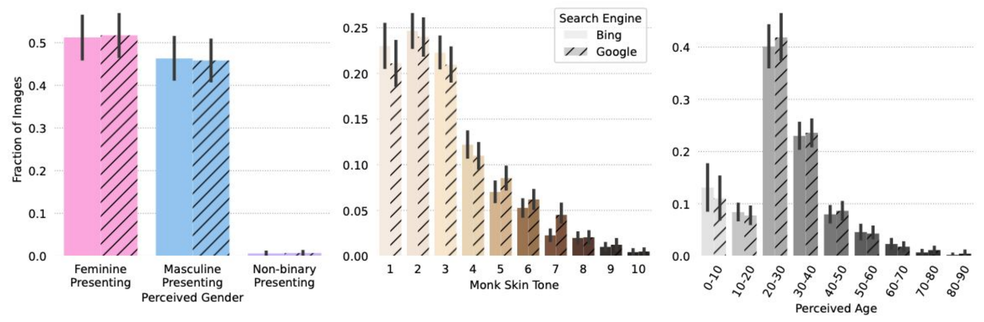
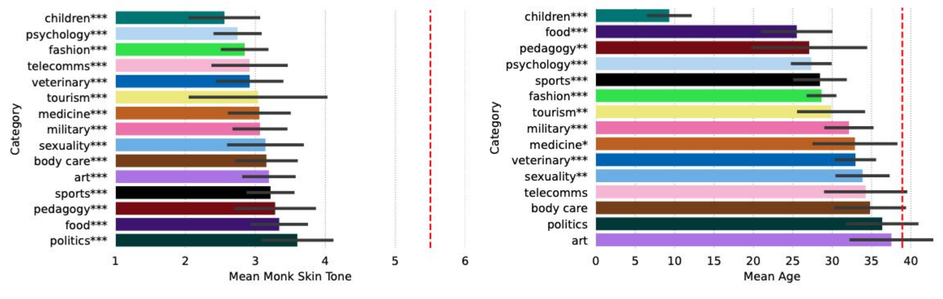


Figure 3: Perceived gender, Monk Skin Tone, and age distributions in image search results. We compute 95% confidence intervals using the percentile bootstrap with 1000 replications over queries.

RQ2: Search results on both Google and Bing are substantially skewed toward lighter-skin tones and away from older people.



(c) Comparing Monk Skin Tone across query categories. **(d) Comparing age across query categories.**

RQ3: We compared representation in each category to a reference baseline. We found that the skews for skin-tone and age were common across most of the top 15 categories.