Dataset Diversity: Measuring and Mitigating Geographical Bias in Image Search and Retrieval

Abhishek Mandal
Insight SFI Research Centre for Data Analytics
School of Computing
Dublin City University
Ireland
abhishek.mandal2@mail.dcu.ie

Susan Leavy
Insight SFI Research Centre for Data Analytics
School of Information and Communication Studies
University College Dublin
Ireland
susan.leavy@ucd.ie

Suzanne Little
Insight SFI Research Centre for Data Analytics
School of Computing
Dublin City University
Ireland
suzanne.little@dcu.ie

ABSTRACT

Many popular visual datasets used to train deep neural networks for computer vision applications, especially for facial analytics, are created by retrieving images from the internet. Search engines are often used to perform this task. However, due to localisation and personalisation of search results by the search engines along with the image indexing method used by these search engines, the resultant images overrepresent the demographics of the region from where they were queried from. As most of the visual datasets are created in western countries, they tend to have a western centric bias and when these datasets are used to train deep neural networks, they tend to inherit these biases. Researchers studying the issue of bias in visual datasets have focused on the racial aspect of these biases. We approach this from a geographical perspective. In this paper, we 1) study how linguistic variations in search queries and geographical variations in the querying region affect the social and cultural aspects of retrieved images focusing on facial analytics, 2) explore how geographical bias in image search and retrieval can cause racial, cultural and stereotypical bias in visual datasets and 3) propose methods to mitigate such biases.

CCS CONCEPTS

• **Computing methodologies** → *Image and video acquisition.*

KEYWORDS

dataset bias; computer vision fairness; visual datasets; image search and retrieval

ACM Reference Format:
Abhishek Mandal, Susan Leavy, and Suzanne Little. 2021. Dataset Diversity: Measuring and Mitigating Geographical Bias in Image Search and Retrieval. In *Proceedings of the 1st Int'l Workshop on Trustworthy AI for Multimedia Computing (Trustworth AI '21), Oct. 24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3475731.3484956

1 INTRODUCTION

Bias in Artificial Intelligence (AI) has the potential to amplify social inequality, systemise oppression and discriminate against marginalised and minority groups. In the domain of computer vision, racial, gender and intersectional biases mirror societal biases and discriminate against those already most marginalised For instance, in the context of computer vision, facial recognition software is least accurate for females and those with darker skin [1]. Video conferencing systems were found to be less effective in creating virtual backgrounds for darker skinned people [4]. Depixelation software was also found to lighten darker skin tones [12]. Such biases undermine public trust in AI and work against principles of fairness and ethics in artificial intelligence.

Biases in computer vision often originate within data used to train deep neural networks. Many of the datasets used for training have been shown to exhibit a 'western centric' bias [5, 13]. These biases can be learned and propagate throughout the machine learning pipeline, leading to the creation of biased computer vision models [3]. For instance, a well established method of compiling datasets to train computer vision models involves retrieving images from the Internet. This technique was used in the computer vision datasets ImageNet [11] and FFHQ [6]. The approach involves either scraping the images from search engines such as Google (as done in case of ImageNet) or scraping from an image hosting site such as Flickr (as done in case of FFHQ). However, this process of image retrieval is not free from bias and can result in biased datasets. In creating image datasets, it is therefore important to audit and evaluate the methods used to retrieve images from the Internet. This paper examines the kinds of biases that may be embedded within datasets using online image search and retrieval methods. The extent to which diversity in datasets is affected by various image search approaches is evaluated and methods to mitigate bias in image datasets are proposed.

Auditing of social bias in visual datasets for faces has relied primarily on two main parameters: race (focusing on skin tone), and gender [1, 3, 5, 13]. We approach this issue from a different perspective – geography. We outline a set of circumstances that leads to a geographical bias, which mimics a combination of societal biases including selection, cultural, confirmation and implicit stereotype bias. We identify issues with techniques used in image retrieval and demonstrate how this contributes to bias in datasets. We then

Trustworth AI '21, October 24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8674-6/21/10.
https://doi.org/10.1145/3475731.3484956

propose methods to evaluate bias and increase levels of diversity in datasets.

Most internet search engines return localised results with respect to the language of a particular query and the IP-address derived location from which it was queried. In this research therefore, we focus on these factors and examine how they affect levels of diversity within the resulting datasets. The effects of language and location on search results is illustrated in Figures 1 and 2 where considerable differences in relation to cultural, racial and gender-based attributes in search results were returned when a query is expressed in different languages and searched from different geographical locations. In examining the extent to which query language and localisation techniques of search engines may result in localised biases in image datasets and how this may be evaluated and mitigated, we address the following questions:

- How does varying the location and language of an online search affect image search and retrieval?
- How is the diversity of visual datasets for facial analytics affected by the image retrieval methodology?

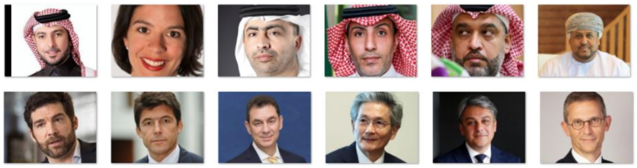

Figure 1: Search results for ‘CEO’. Arabic-UAE (top), English-UK (bottom)

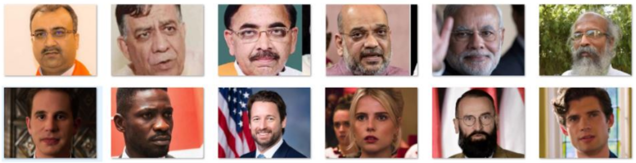

Figure 2: Search results for ‘Politician’. Hindi-India (top), English-USA (bottom)

2 RELATED WORK

Social biases in multimedia computing pertaining to race, gender and its intersections have been studied in great detail in the recent years. In computer vision, these biases are particularly apparent in facial recognition technology. Buolamwini and Gebru [1] audited commercial facial recognition systems and found that facial classifiers perform better on male faces and those with lighter skin tone. This shows the presence presence of racial, gender and intersectional bias in them with white males getting the best performance and black females the worst. Similar biases exist in other technologies involving computer vision such as digital cameras and AI judges for beauty contests [9]. Mehrabi et al. [9] reviewed the types of biases present in AI systems including historical, representational, evaluation and sampling bias among many others. Most of these biases manifest themselves as gender and racial bias. As such, in facial recognition, most of the research focuses on two aspects: skin tone and gender.

Celis and Keswani [3] studied the presence of social bias in image retrieval and found that search queries related to people tend to return images that reflect the social, cultural and demographic attributes of the region. This leads to over-representation of the demographic of the region from where the images are being queried. Machine learning algorithms tend to be more prone to implicit biases [3]. As such, they tend to amplify biases present in datasets. Any information that is not explicitly fed to the models during training, simply does not exist for it. For example, if all the images labelled as ‘CEO’ are of white men, the model will not associate women or men of colour with that particular term. Therefore, having diversity in training datasets is imperative to countering and mitigating implicit biases in machine learning models [3, 13]. Building on this work we approach the issue of bias in facial recognition datasets from the perspective of geography and devise measures that evaluate and accordingly increase diversity.

2.1 Biases in Visual Datasets

Karkkainen and Joo [5] analysed various popular visual datasets. Their study included Labeled Faces in the Wild, CelebA, COCO, IMDB-WIKI, VGG2, DIF and UTK datasets. Their findings show that the majority of the faces in the datasets, from ~45% in the UTK dataset to ~88% in the LFWA dataset, are white. Most of these datasets are either created or drawn from sources in western countries [2, 6–8, 11]. The sources include web scraping using search engines, social networking sites, news and other media. As a result, they have a high representation of attributes associated with Western societies such as faces with lighter skin tone and western clothing, leaving the datasets heavily biased with an under-representation of non-western regions such as Africa or West Asia. Wang et al. [13] studied the geographical distribution of images in OpenImages and ImageNet and found them to be Europe and North America centric, with the USA being highly over-represented and Africa being severely under-represented. When these datasets are used to train deep learning models such biases can be propagated within the learning models and amplified within AI systems [3].

Bias in datasets may be of many types. Mehrabi et al. [9] discussed the various types of social and cognitive biases in data and their effects on downstream machine learning pipelines. Many of the biases (such as measurement or behavioral bias) cannot be used to evaluate visual datasets due to its nature. Celis et al. [3] and Wang et al. [13] analysed a few causes of bias in visual datasets arising due to historical and cultural reasons. Understanding the range of biases that may be embedded in a dataset is key to identifying potential sources of bias and mitigating them. The following presents an overview of the predominant kinds of biases in image datasets:

- ***Historical Bias:*** Mehrabi et al. [9] defines historical bias as preexisting bias present in society that seeps into the data, even after perfect sampling. An example of such bias would be that image search results for the term ‘CEO’ would be mostly men. This reflects that, both historically and currently, the majority of CEOs positions were held by men [3].

- ***Cultural Bias:*** Cultural bias may be defined as certain cultural features and attributes, which are associated with a certain group or sub-group of people. For example, analysis of labels in popular visual datasets revealed that images of women are associated with labels such as 'kitchen', 'indoor', 'food' and 'appliances' while those of men are associated with labels such as 'sports', 'vehicle' and 'outdoor' [13].
- ***Representation Bias:*** This type of bias is often caused due to poor sampling or less diversity in the source. For example, Labelled Faces in the Wild (LFW) dataset was curated from news articles in the USA [7]. This is reflected in the dataset with about 88% images being of white people [5].

2.2 Search Engine Localisation as Cause of Bias

Search engines such as Google and Bing generally aim to provide search results that are relevant to the geographical location according to query settings or derived from the originating IP address. Retrieving images in this way returns images with local characteristics. Although a good technique for providing users with locally relevant results, creating datasets in this way can limit the resulting diversity as shown in section 2.1.

Researchers have attempted to address this issue by adding location based keywords along with the queries. For example datasets used query terms such as "Asian Boy" for image retrieval but this can lead to stereotypical bias in the results (e.g. the images being mostly of East Asian people with little or no representation of people from other regions of Asia) [5, 11]. Given that search engines index images based on keywords from web pages, queries in English will return images from an English language website [3]. For example, for a query such as 'Farmers in Thailand', the images returned are mainly from English language websites. Search results for 'Egypt' and 'Paris' similarly return images of pyramids and the Eiffel Tower. The images reflect concepts garnered from websites containing content in English and are thus more likely to reflect a Western viewpoint. Websites written in the main language of a particular country are excluded when the queries are not in the local language [13]. Attempts to mitigate this issue when creating datasets have been taken by translating search queries into different languages. ImageNet, for example, translated the queries into Chinese, Spanish, Dutch and Italian, along with English. However, apart from Chinese, all other languages are European and thus, have a Western centric bias [11]. Addressing this bias in image search and retrieval as a method of generating image datasets is the main focus of this paper.

3 DEFINING 'GEOGRAPHICAL BIAS'

A combination of the localisation and personalisation of search engine results along with the issue of implicit bias in machine learning algorithms, gives rise to a kind of bias that we have defined as geographical bias. In defining the concept of geographical bias, we build on ideas from social identity theory and social psychology and apply them in the context of information retrieval. We define *geographical bias* as a type of selection bias which is at the intersection of cultural bias, confirmation bias and implicit stereotype bias. The following outlines contributing factors and impacts of geographical bias and how they are caused by localisation and personalisation in online image search:

Contributing Factor → Selection bias: Localisation and personalisation features of a search engine along with language of the query can result in a a biased search query. This paper identifies selection bias in the search query as a central cause of bias in many image datasets.

Outcome → Cultural bias: Limited variation in the language of the search terms along with localisation and personalisation of the search engine results in biased image datasets.

Societal Impact → Confirmation and implicit stereotype bias: Due to localisation and personalisation features of the search engines, the images retrieved confirms or supports prior concepts and beliefs concerning the local population. For example, the stereotypical concept of a CEO being a 'white middle aged male wearing a business suit' in western society could be reinforced. All the above factors contribute to an unintentional bias learned through repetition.

4 CREATING THE DATASET

To test our hypothesis, we collected a dataset of human faces from a wide geographical range. To include as much variation as possible, we divided the world into nine regions, each having a language (the *lingua franca* of that particular region) and one or two countries from the region. The countries chosen are generally the more populous countries of that region. The choice of regions is inspired from the racial groupings by Karkkainen and Joo [5] based on the US Census Bureau. We added further categories to increase cultural diversity.

4.1 Language-Location Pair and Query Selection

We divide the world into the following nine regions: East Asia (EA), South Asia (SA), South East Asia (SEA), North Africa and West Asia (NAWA) , Sub Saharan Africa (SSA), West Europe (WE), East Europe (EE), North America (NA) and Latin America (LA). We selected the most widely spoken language in each region as the query language and the most or second most populous country in that region as the querying location resulting in nine sets of language-location pairs. Note that location here refers to the region and not the individual country.

Five queries were selected as query terms: 'CEO', 'Engineer', 'Politician', 'Nurse' and 'School Teacher'. All these queries represent identities based on occupation. 'CEO' and 'Engineer' are traditionally considered male dominated fields whereas 'Nurse' and 'School Teacher' are female dominated. 'Politician' is considered globally more gender balanced. We translated these queries into different languages using Google translate, as per the language-location pairs. See [1] for the full list of language-location pairs and the corresponding queries in that language.

4.2 Image Collection and Face Cropping

We used Google advanced image search for querying the images and used a Virtual Private Network (VPN) as well as the 'region' option

[1] https://github.com/aibhishek/Geographical_Bias/blob/main/Query_Terms.pdf

in image search to specify the origin country for each query. Each language-location pair corresponding to one region has five queries in that particular language. For each query, 150 images were scraped. For language-location pairs with two countries, 75 images were queried from each. For queries in Spanish where the noun changes with gender, we used each form of the noun and included the search with 'any of these words' feature. For example, search results for 'politician' would include results for either 'politica' or 'politico'. The 150 images were the top results returned by Google's algorithm. To detect the faces in the images and crop them, we employed a face detection and cropping algorithm called Autocrop[2]. We randomly select 20 faces for each query. A total of 900 faces were collected for all the regions. This collection of 20 faces, corresponding to one query for one language-location pair, is referred to as an image set.

Each face image has a dimension of 650x500 pixels, with the face covering 80% of the entire image. The resolution was determined experimentally. The background, which covered the remaining 20% of the image, along with attributes such as any visible clothing and headwear were retained. Headwear, clothing and background constitute important cultural aspects of a person's identity. As our aim was to reduce stereotypical bias and increase diversity, it was important to include these details. Also certain professions have uniforms that may include headwear such as engineer's safety hat and nurse's head covering and mask. Finally, we randomly select two images from each language-location pair and create a *Diverse Dataset* that we use as a benchmark against which to compare the diversity of other popular datasets.

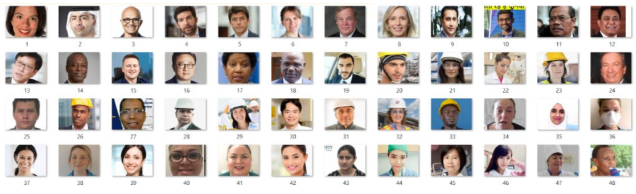

Figure 3: Sample of cropped faces.

5 IMAGE SIMILARITY

We calculate image similarity score following an approach used by Celis and Keswani [3]. Image Similarity Score (ISS) compares how similar two images are based on the features extracted by a pre-trained convolutional neural network (CNN). We use VGG16, a 16 layer deep CNN, trained on the Imagenet dataset, created by the Visual Geometry Group at the University of Oxford. The feature extracting layers of the VGG16 were used to extract the features from the image. Dimensions of the extracted features were reduced using principal component analysis. For two images I_1 and I_2, with extracted features v_1 and v_2 respectively, we calculate image similarity as:

$$sim(I_1, I_2) = 1 - \frac{v_1 \cdot v_2}{||v_1||_2 \cdot ||v_2||_2}$$

$$sim(I_1, I_2) \in [0, 2]$$

The image similarity score is 1 - cosine similarity of the two feature vectors. As such, for two copies of the same image, the cosine similarity will be 1 ($\theta = 0$), and therefore the image similarity score is 0. For two decorrelated images, the value of cosine similarity will be 0 ($\theta = \frac{\pi}{2}$), and therefore the value of image similarity score will be 1. Similarly, for two visually opposite images, the value of cosine similarity will be -1 ($\theta = \pi$), and therefore the value of image similarity score will be 2. The higher the image similarity score therefore, the more different the image is.

5.1 Visual Similarity

Visual similarity in this paper refers to how similar two images of people look [3]. Features such as eyes, lips and noses follow a pattern whereby the eyes for example, are always above the nose. These patterns are picked up by Convolutional Neural Networks (CNN). However, they differ in terms of skin colour, facial structure and clothing and constitute a different set of features which vary as per region, ethnicity and gender [3]. These patterns are also picked up by CNNs. If training datasets consist of faces from a particular region or ethnicity or gender, the CNNs start to identify those social attributes with humans. And when those CNNs are used on faces having attributes different from those they were trained with, they fail to identify those faces correctly [1, 3, 5, 13]. Figure 4 shows a heatmap of similarity scores of images of people from different regions. We can see that people from similar regions have a lower similarity score indicating that they have higher visual similarity. The similarity score of two identical images is zero.

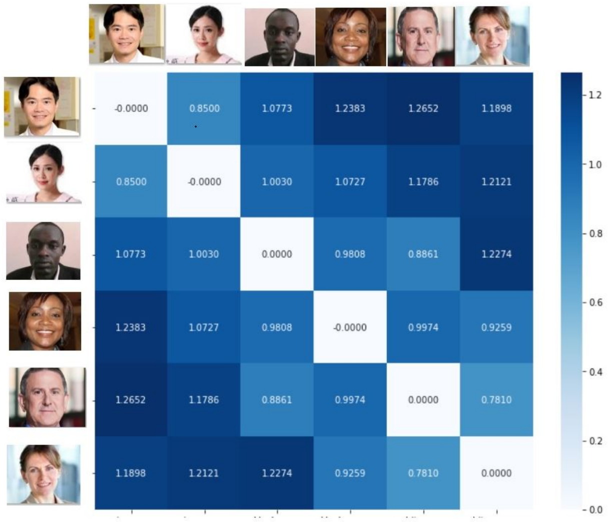

Figure 4: Heatmap of image similarity.

5.2 Image Similarity Scores

To test our hypothesis, we created two variations of the image similarity scores: intra and cross. Image similarity *intra* is the mean of image similarity score of every image with all other images in the image set. Image similarity *cross* is the mean of all of the image similarity scores between the images across all the given image

[2] https://github.com/leblancfg/autocrop

sets. Image similarity intra will compare one image set while cross will assess multiple image sets.

Algorithm 1 Algorithm for Calculating Image Similarity Score - Intra

```
Input: List of images (Image Set) $S$
for each $I \in S$ do
    convert $I$ to array
    process the array and extract features
    reshape extracted feature array to 20x15x512
    apply PCA transform
    reshape to dimensions 1x400
    append reshaped feature vector array to list $V$
copy $V \rightarrow V'$
for each $F \in V$ do
    for each $F' \in V'$ do
        Calculate $sim(F, F')$
        append similarity score to list $L$
return mean($L$)
```

Algorithm 2 Algorithm for Calculating Image Similarity Score - Cross

```
Input: List of list of images (Image Set) $S$
for each $S' \in S$ do
    for each $I \in S'$ do
        extract features
        reshape extracted feature array to 20x15x512
        apply PCA transform
        reshape to dimensions 1x400
        append reshaped feature vector array to list $V'$
    add $V'$ to stack $V$
while $V \neq \{\}$ do
    pop item from $V$ and assign to $A$
    for each $A' \in V$ do
        for each $F \in A$ do
            for each $F' \in A'$ do
                Calculate $sim(F, F')$
                append similarity score to list $L'$
        append $mean(L')$ to list $L$
return mean($L$)
```

6 EXPERIMENT

We began by arranging the image sets according to all the language-location pairs for each query term. Each query term (5 in total), consists of 9 language-location pairs. Each language-location pair consists of one image set i.e., a list of 20 images. Fig. 5 shows this arrangement for the query term 'CEO' and language-location pair 'Arabic-NAWA'. The language-location pairs for all the query terms remain the same[3].

Image similarity score – Intra ($\text{ISS}_{\text{intra}}$) is calculated by taking the mean of sim(I,I') of all the images in a single image set. The mean of $\text{ISS}_{\text{intra}}$ for all the language-location pairs for a given query gives $\text{ISS}_{\text{intra}}$ for that query term.

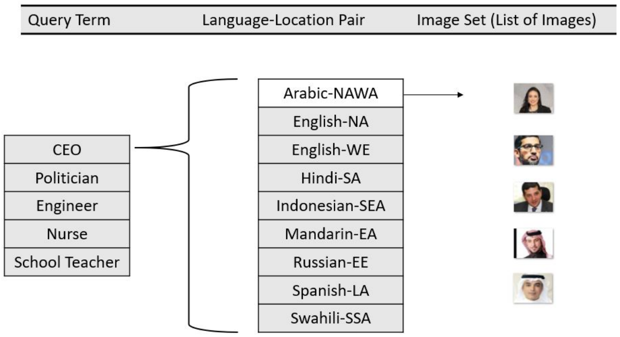

Figure 5: Arrangement of Images according to Query Term and Language-Location Pair.

Query	$\text{ISS}_{\text{intra}}$	$\text{ISS}_{\text{cross}}$
CEO	0.9644	**0.9846**
Engineer	0.9925	**0.9939**
Nurse	0.9862	**0.99**
Politician	0.9724	**0.9836**
School Teacher	0.9860	**0.9904**
Mean Value	0.9803	**0.9885**

Table 1: $\text{ISS}_{\text{intra}}$ and $\text{ISS}_{\text{cross}}$ scores for all the query terms. Higher values have been highlighted in bold.

Image similarity score – Cross ($\text{ISS}_{\text{cross}}$) is calculated by taking the mean of sim(I,I') of each image in every image set with all the images of all other image sets for a particular query term. We calculate $\text{ISS}_{\text{intra}}$ and $\text{ISS}_{\text{cross}}$ for all the query terms and tabulate the results.

6.1 Calculating Diversity in Visual Datasets

We use $\text{ISS}_{\text{intra}}$ as a way to measure diversity in some popular face datasets. For this, we chose Flickr Faces HQ (FFHQ), WIKI, IMDB, Labelled Faces in the Wild (LFW), UTK Faces and our own Diverse Dataset. To calculate $\text{ISS}_{\text{intra}}$, 100 images were randomly sampled from each of the datasets (except Diverse Dataset, as it had only 90 images). We then calculated $\text{ISS}_{\text{intra}}$, and tabulated the results (Table 1 and 2).

7 FINDINGS AND DISCUSSION

The findings of this study demonstrate how varying the language and location of online image search queries can increase the diversity of the resulting visual dataset. In comparing the diversity of datasets in our study we found that the values of $\text{ISS}_{\text{cross}}$ are higher than $\text{ISS}_{\text{intra}}$ for all the query terms (see Table 1). Consequently, the mean value of $\text{ISS}_{\text{cross}}$ is also higher. The $\text{ISS}_{\text{cross}}$ values were higher than the individual $\text{ISS}_{\text{intra}}$ scores for 66.66% of cases (30 out of 45 language-location pairs). Appendix A contains the individual scores and calculations for each query and language location pair. These quantitative measurements support our visual observation

[3] https://github.com/aibhishek/Geographical_Bias/blob/main/Query_Terms.pdf

Dataset	$\text{ISS}_{\text{intra}}$
FFHQ [6]	0.9940
Diverse Dataset	0.9895
WIKI [10]	0.9786
IMDB [10]	0.9661
LFW [7]	0.9536
UTK [14]	0.9418

Table 2: $\text{ISS}_{\text{intra}}$ of Datasets.

outlined in section 1 that the image search results obtained by varying the language and location of the search queries increases the visual diversity of the results.

The quantitative results (Figure 6) show that $\text{ISS}_{\text{intra}}$ has a wider spread than $\text{ISS}_{\text{cross}}$, which are more concentrated around one value (~1). At sim(I,I') = 1, the images are decorrelated, or highly visually dissimilar. This means that the images are least correlated at this point.

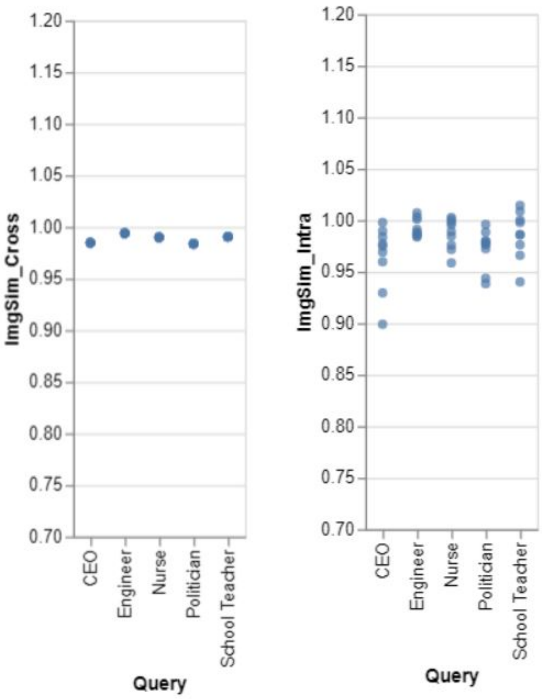

Figure 6: Distribution of $\text{ISS}_{\text{intra}}$ and $\text{ISS}_{\text{cross}}$.

The 'Diverse Dataset' has higher $\text{ISS}_{\text{intra}}$ than all the other datasets except FFHQ (see Table 2 and Figure 7). Thus we can conclude that using a diverse querying methodology (by varying the language and location of search queries) in image retrieval for creating visual datasets can lead to better visual diversity of datasets and help mitigate geographical bias inherited from the search systems.

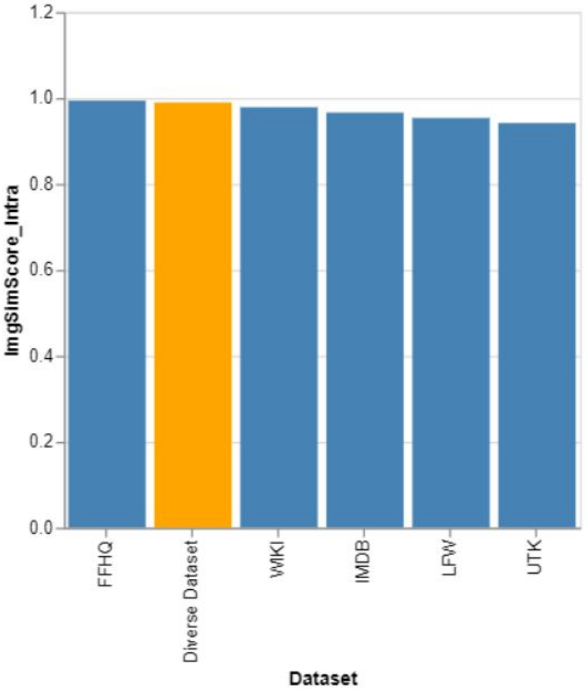

Figure 7: Distribution of $\text{ISS}_{\text{intra}}$ of Datasets (Bar Chart).

8 CONCLUSION

This study presents a methodology for increasing the diversity of datasets and mitigating the effects of geographical bias by increasing the variations in the language of the query terms and the location of the search engines. We demonstrated how current methods of compiling visual datasets using online image retrieval can introduce a type of bias that is unique to this process and reflects a variety of social biases. In order to study and understand this bias more thoroughly, we defined geographical bias, based on the parameters that introduce it: language and location of search engine queries. This type of bias manifests in different forms, throughout the machine learning pipeline, as racial, cultural and stereotypical bias.

In this study, we created a dataset that can reflect the visual diversity of human faces at a global level. However, human faces are extremely diverse with respect to physiological attributes such as face structure and skin colour as well as cultural attributes such as headwear rendering it difficult to capture a full range of diversity in a single dataset. Our goal in this research is therefore to provide a framework for increasing visual diversity in training datasets for training deep neural networks. This methodology can be scaled as per the cultural, racial and ethnic variations of the demographics, on which the final algorithm is targeted.

ACKNOWLEDGMENTS

Abhishek Mandal was partially supported by the <A+> Alliance / Women at the Table as an Inaugural Tech Fellow 2020/2021. This publication has emanated from research supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

REFERENCES

[1] Joy Buolamwini and Timnit Gebru. [n.d.]. Gender Shades. http://gendershades.org/overview.html

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2018. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. https://doi.org/10.1109/fg.2018.00020

[3] L. Elisa Celis and Vijay Keswani. 2020. Implicit Diversity in Image Summarization. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–28. https://doi.org/10.1145/3415210

[4] Megan Rose Dickey. 2020. Twitter and Zoom's algorithmic bias issues. https://techcrunch.com/2020/09/21/twitter-and-zoom-algorithmic-bias-issues

[5] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1548–1558.

[6] Tero Karras, Samuli Laine, and Timo Aila. [n.d.]. NVlabs/ffhq-dataset. https://github.com/NVlabs/ffhq-dataset

[7] Erik Learned-Miller, Gary B. Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. 2016. Labeled Faces in the Wild: A Survey. In *Advances in Face Detection and Facial Image Analysis*. Springer International Publishing, 189–248. https://doi.org/10.1007/978-3-319-25958-1_8

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*. Springer International Publishing, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[9] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[10] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. DEX: Deep EXpectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (April 2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[12] James Vincent. 2020. What a machine learning tool that turns Obama white can (and can't) tell us about AI bias. https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

[13] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. 2020. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. In *Computer Vision – ECCV 2020*. 733–751.

[14] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4352–4360. https://doi.org/10.1109/CVPR.2017.463

A IMAGE SIMILARITY SCORES FOR ALL QUERIES

Query	Language Location Pair	ISS_intra	ISS_cross
CEO	Arabic-West Asia & North Africa	0.899012	0.984683
	English-North America	0.968974	
	English-West Europe	0.929469	
	Hindi-South Asia	0.997845	
	Indonesian-SE Asia	0.983675	
	Mandarin-East Asia	0.989452	
	Russian-East Europe	0.959661	
	Spanish-Latin America	0.974743	
	Swahili-Sub Saharan Africa	0.977119	
Engineer	Arabic-West Asia & North Africa	0.98639	0.993904
	English-North America	0.988344	
	English-West Europe	1.000911	
	Hindi-South Asia	1.003149	
	Indonesian-SE Asia	0.987191	
	Mandarin-East Asia	0.991146	
	Russian-East Europe	1.007155	
	Spanish-Latin America	0.984955	
	Swahili-Sub Saharan Africa	0.983727	
Nurse	Arabic-West Asia & North Africa	1.002607	0.989952
	English-North America	0.971564	
	English-West Europe	0.99561	
	Hindi-South Asia	0.984535	
	Indonesian-SE Asia	0.975914	
	Mandarin-East Asia	0.98904	
	Russian-East Europe	0.997979	
	Spanish-Latin America	1.000587	
	Swahili-Sub Saharan Africa	0.958532	
Politician	Arabic-West Asia & North Africa	0.977348	0.983637
	English-North America	0.995927	
	English-West Europe	0.979358	
	Hindi-South Asia	0.979915	
	Indonesian-SE Asia	0.972307	
	Mandarin-East Asia	0.976251	
	Russian-East Europe	0.93835	
	Spanish-Latin America	0.988452	
	Swahili-Sub Saharan Africa	0.943626	
School Teacher	Arabic-West Asia & North Africa	1.014298	0.990403
	English-North America	0.997715	
	English-West Europe	0.940142	
	Hindi-South Asia	1.000047	
	Indonesian-SE Asia	0.985991	
	Mandarin-East Asia	1.00862	
	Russian-East Europe	0.976169	
	Spanish-Latin America	0.965902	
	Swahili-Sub Saharan Africa	0.985919	