FYP Research - A survey on bias in visual datasets

[A survey on bias in visual datasets](https://www-sciencedirect-com.ejournals.um.edu.mt/science/article/pii/S1077314222001308?via%3Dihub)

# **1. Introduction**

The study aims to address the concern of bias in datasets by

1. Identifying various forms of bias in visual datasets.
2. Reviewing methods for detecting and measuring bias.
3. Discussing efforts to create datasets with bias awareness.

Algorithms may be responsible for the amplification of pre-existing biases in the training data, issues in the quality of the data itself could contribute significantly to the development of discriminatory AI applications.

Two ways in which bias is encoded in the data were identified:

1. Correlations and causal influences among the protected attributes and other features
2. The lack of representation of protected groups in the data
   * It was also noted that biases manifest in ways that are specific to the data type.

# **2.** [**Manifestation of bias in visual data**](https://www-sciencedirect-com.ejournals.um.edu.mt/science/article/pii/S1077314222001308?via%3Dihub#sec2)

Types of bias that pertain to the capture and collection of visual data:

1. **Selection bias** - The way we pick which pictures or images to include can create unfair differences or connections. This can lead to mistakes in understanding or results. For example, if we're studying faces and only include certain types of people, we might not get a complete picture of how well our system works for everyone. So, we need to be careful when choosing what images to use in our datasets to avoid this bias. **(Simplified Explanation via ChatGPT)**
2. **Capture/Framing bias** - When we "frame" something, we choose certain parts of what we're talking about and make them more noticeable in our communication. This helps shape how people see the issue and what they think about it. This concept of framing isn't just for words – it also applies to pictures. In visual studies, framing means picking a particular view, scene, or angle when creating or editing an image. These definitions show that framing bias has two parts. First, the way an image is put together can send different messages. Second, how an image is taken or edited can also introduce bias. So, when we talk about framing bias, we mean any differences or connections in an image that make people think differently, and these differences might come from how the image was put together.
3. **Label bias** - In supervised learning, we need labelled data to teach the computer. The accuracy of these labels is crucial but can be challenging because today's datasets are complex and vast. Label bias happens when the labels given to the computer are different from the actual truth. For example, a face recognition dataset might have labels that are not very accurate compared to human annotations. Sometimes bias also comes from how things are labelled. For instance, different people might call the same thing by different names. This can be a big problem when dealing with things related to people like race or gender. So, label bias means mistakes in labelling data, either because the labels are wrong compared to the truth or because they use unclear or inappropriate categories.
4. **Negative set bias** - The labelling does not reflect entirely the population of the negative class (say non-white in a binary feature [white people/non-white people]). Negative class bias is being considered as an instance of selection and label bias.

A close-up of several images of different types of bias

Description automatically generated

* During the capturing section of the data selection & framing bias can be introduced
* Dissemination of visual content suffers from both selection and framing bias.
* Data collection for the sake of creating a dataset can result in selection & label bias.
* Algorithms also raise issues as biased generative models can generate images which can result in further bias in the Visual Content Life Cycle.

# **3. Bias discovery and quantification in visual datasets**

Section 3 of the paper aims to answer the following "Given a visual dataset, is it possible to discover/quantify what types of bias it exhibits?"

The strategies used to discover bias:

**Reduction to tabular data** (Imp for our case)

1. These rely on the attributes and labels attached to or extracted from the visual data and try to measure bias as if it were a tabular dataset.
2. The features for the tabular description can be extracted either **directly from the images** (using, for example, some image recognition tools) or **indirectly from some accompanying image description/annotation** or both. This is therefore prone to errors and biases.
3. The biases that exist in the original images (selection, framing, or label bias) might be reflected and even amplified in the tabular representation due to the bias-prone feature extraction process.
4. Bias may also exist due to the labelling process and the automatic feature extraction. The impact of such additional sources on the results is typically omitted.

**Biased image representations**

1. These rely on lower-dimensional representations of the data to discover bias.

**Cross-dataset bias detection**

1. These assess bias by comparing different datasets, trying to discover some sort of “signature” due to the data collection process.

**Other methods**

1. Different methods that could not fit any of the above categories.

**Count/demographic parity methods (Reduction to tabular data)**

**Method:**

* Proposed a method to audit biases related to gender and age in the ImageNet dataset.
* Applied a face detection algorithm to two subsets of ImageNet:
  + ILSVRC training set (Russakovsky et al., 2015).
  + Person category of ImageNet (Deng et al., 2009).
* Employed age and gender recognition models on these subsets.
* Calculated dataset distribution across age and gender categories.

**Findings:**

* Discovered a gender prevalence: Men constituted 58.48% of the dataset.
* Limited representation of people in specific age groups (1.71%).
* Analysed percentage of men and women across categories.
* Identified class imbalances in ImageNet subsets:
  + Example: In the ILSVRC subset, 89.33% of "bulletproof vest" images labelled as men, 82.81% of "lipstick" images labelled as women.
* Method also offers insight into selection bias and framing of protected attributes with appropriate dataset labelling.

**Limitation:**

* Reliance on unbiased gender and age recognition models.
* Noted that their gender recognition model was biased, affecting analysis reliability.

**Method:**

* Conducted analysis on the person category of ImageNet to address selection & label bias.
* Addressed label bias:
* Annotators evaluated labels for offensiveness or sensitivity (e.g., racial/sexual slurs).
* Removed labels not corresponding to visual categories (e.g., non-visual labels like "philanthropist").
* Annotators we asked to categorized images based on gender, age, and skin colour. This was aimed at assessing data balance within these categories and addressing selection bias.

**Findings:**

* Demographic analysis revealed under-representation of women and dark-skinned individuals in the remaining non-offensive/sensitive and visual categories.
* Some categories aligned with stereotypes, Hence the potential addressing of framing bias due to alignment with stereotypes.
* Example: 66.4% of people in "rapper" category were dark-skinned.

**Validation:**

* The annotation process was validated by measuring the annotators’ agreement on a small, controlled set of images.

Further methods were also used which included:

* Measuring the correlation between a protected attribute and the occurrences of various objects/actions. (Example a gender with a specific cooking tool or outdoor sport)
* Geographic bias was also measured via the use of a simple count, it was found that a great majority of images of which the geographic provenance is known came from the USA or [Western European countries](https://www-sciencedirect-com.ejournals.um.edu.mt/topics/computer-science/western-european-country), resulting in highly [imbalanced data](https://www-sciencedirect-com.ejournals.um.edu.mt/topics/computer-science/imbalanced-data).
* [Buolamwini and Gebru (2018)](https://www-sciencedirect-com.ejournals.um.edu.mt/science/article/pii/S1077314222001308?via%3Dihub#b13) constructed a benchmark dataset for gender classification. For testing discrimination in gender [classification models](https://www-sciencedirect-com.ejournals.um.edu.mt/topics/computer-science/classification-models), their dataset is balanced according to the distribution of both gender and Fitzpatrick skin type as they noticed that the error rates of classification models tended to be higher at the intersection of those categories (e.g., black women) because of the use of imbalanced training data. Hence, while they quantify bias by simply counting the instances with certain protected attributes, the novelty of their work is that they considered multiple protected attributes at a time.

**Information theoretical (Reduction to tabular data)**

* Using four measures to construct a balanced dataset of faces. Two measures of diversity Shannon entropy & Simpson index and two measures of evenness using Shannon and Simpson as well.
* Proposed to use (conditional) Shannon entropy for discovering *framing bias* in emotion recognition datasets.
* Using the precise definition of bias: a dataset contains bias when **I(X,Y)≫0** where **X** is the protected attribute, **Y** is the output variable and **I(X,Y) ≔ H(X) − H(X|Y)** is the mutual information between those [random variables](https://www-sciencedirect-com.ejournals.um.edu.mt/topics/engineering/random-variable-xi). Kim et al. proposed to minimise such mutual information during training so that the model forgets the biases and generalises well.

**Note:** The “**other**” section contains methods which further expanded on the above methods or specified on specific types of data/datasets.

### **Biased image representations**

* The distance and geometric relations between images were measured after passing them through a neural network to convert them into lower dimensional forms.
* The method involves measuring diversity in a face dataset by analysing pairwise L1-distances of images after embedding them into a lower-dimensional space using a neural network pre-trained on a different face dataset, where a skewed distribution towards higher distances signifies higher data diversity, but potential bias in the embedding can lead to misleading results.
* Bias was measured by looking at associations between [semantic concepts](https://www-sciencedirect-com.ejournals.um.edu.mt/topics/computer-science/semantic-concept) (for example, man-career and woman-family) by measuring the [cosine similarity](https://www-sciencedirect-com.ejournals.um.edu.mt/topics/computer-science/cosine-similarity) among vectors in the latent space computed by applying the models to controlled samples of images that resemble those visual concepts.
* A method for assessing *algorithmic* biases in image classifiers was developed via computing the [causal relationship](https://www-sciencedirect-com.ejournals.um.edu.mt/topics/computer-science/causal-relationship) between the protected attribute and the output of a classifier.

### **Cross-dataset bias detection**

* Bias in object detection datasets were tested by answering the following question: “How well does a typical object detector trained on one dataset generalise when tested on a representative set of other datasets, compared with its performance on the native test set?” The assumption here is that if the performance on the native test set is much higher it means the datasets exhibit some *bias* that is learned by the object detector.

### **Model-based bias detection**

* A simple method for addressing bias in object recognition datasets. They cropped a small central sub-image from each image in the original dataset. These cropped pictures were so small that humans could not recognise the objects in the pictures. Hence, if the attained performance of a model trained on such images was better than pure chance, the data contained distinguishable features spuriously correlated with the object categories. Then the data showed some kind of selection or framing (capture) bias.

### **Human-based bias detection**

**Approach:**

* Proposed non-automated method for bias detection.
* Consists of a three-step crowdsourcing workflow.

**Steps:**

1. Similarity Description:

* Workers presented with a batch of images.
* Asked to describe similarities among images using question-answer pairs.
* Example: If all images show white airplanes (selection bias), worker labels batch with question "What colour are the airplanes in the images?" and answer "White."

2. Confirmation of Biases:

* Each worker answers questions from the first step using different batches of images.
* Aims to confirm presence of biases identified in step one.

3. Statement Evaluation:

* Workers evaluate statements for real-world truth or potential biases.
* Judged based on common sense knowledge and subjective belief.
* Relies heavily on workers' backgrounds and biases.

**Note:** The final step's subjectivity emphasizes workers' individual perspectives, influencing bias assessment outcomes.

## **Discussion**

The discussion section outlines general use cases and issue with the aforementioned methods:

Reduction to tabular data methods

* Reasonable and effective way of discovering bias.
* Methods used are relatively simplistic.
  + Look for balance in the protected attribute.
  + Compare the distribution with respect to other features.
* Methods heavily rely on labels that are either attached to the data or automatically extracted.
  + Labelling bias affects such discovery methods.

Image representation methods

* These methods better capture the complexity of visual content,
* Such methods are necessarily influenced by both the models used to compute the representation and the metric used to compute distances in it.
* Harder to apply since they need some kind of supervision for computing the latent space.

Cross dataset detection methods

* Only applicable when several comparable datasets are available.
* They might help to unveil the presence of bias, but without further inspections, they cannot reveal what kind of bias it is.
* Useful to get an idea of the existence of some biases in a visual dataset, they are of little or no use if we want to discover bias within the dataset.

# **4. Bias-aware visual data collection**

This section was skipped as it revolved around the construction of datasets which don’t involve bias which was beyond the scope of our FYP.

# **5. Conclusions and research outlook**

This section goes over what was already discussed throughout the paper.

FYP Research - Unbiased look at dataset bias

[Unbiased look at dataset bias](https://ieeexplore-ieee-org.ejournals.um.edu.mt/document/5995347)

# **1. Introduction**

With the aim being to study why certain trained individuals could quite accurately determine to which dataset an image belonged it was chosen to randomly sample 1000 images from the training portions of each of the 12 datasets and train a 12-way linear SVM classifier. Classifier performance was measured for 4 image descriptors 32x32 thumbnail, both grayscale and colour, gist, and bag of HOG visual words.

This led to the obvious conclusion that each dataset possesses a unique signature. This signature was visualised by looking at the most discriminable images within each dataset, i.e., the images placed furthest from the decision boundary by the SVM. Looking at the images closest to the decision boundary would show how one dataset can impersonate another.

This leads to the idea that each dataset has some in built bias innately from the goal the dataset is aiming towards. However, bias isn’t exclusive to this as the classifier gave a 61% performance in determining to which dataset an image belonged to when only images of cars were taken into consideration.

# **2. Prologue: The Promise and Perils of Visual Datasets**

This section discusses how researchers concerning datasets has been more focused on winning competitions and trying to be better than the previous dataset that instead of helping us train models that work in the real open world, they have become closed worlds unto themselves. This is due to representing the world in a skewed manner due in part to the images they consider.

# **2.1. The Rise of the Modern Dataset**

This section goes over a variety of datasets any why they were created in terms of their history. In addition, we as a community tend to reject the current datasets due to their perceived biases. Yet time and again, we create new datasets that turn out to suffer from much the same biases, though differently manifested. What seems missing, then, is a clear understanding of the types and sources of bias, without which, we are doomed to repeat our mistakes.

# **3. Measuring Dataset Bias**

For the purposes of object recognition, most existing datasets assume roughly the same general task given the typical visual environments encountered by people, to detect commonly occurring objects. Using that as the definition of our visual world, can we evaluate how well does a particular dataset represent it? To correctly measure a dataset’s bias would require comparing it to the real visual world, which would have to be in form of a dataset. This however isn’t a viable option as it could be biased. So, here we will settle for a few standard checks, a diagnostic of dataset health if you will.

One of the problems we face is cross dataset generalisation wherein we train on dataset A but testing on dataset B. Given that we aim to represent the real world doing such a thin should be easy however it is not. Although methods do exist which transfer a model of one dataset onto another this shouldn’t be required given that they aim represent the same domain (real world).

# **3.1. Cross-dataset generalization**

There are virtually no papers demonstrating cross-dataset generalization. To answer the question “how well does a typical object detector trained on one dataset generalize when tested on a representative set of other datasets, compared with its performances on the “native” test set?”. To answer this question a set of six representative datasets which are in active research use today, and have some annotated objects in common:

* SUN09
* LabelMe
* PASCAL VOC 2007
* ImageNet
* Caltech-101
* MSRC

Given that the datasets have objects labelled with bounding boxes, two types of testing were carried out:

* classification – find all images containing the desired object.
* detection – in all images, find all bounding boxes containing the desired object.

The object detection task was carried out using a standard, off-the-shelf approach of Dalal&Triggs (HOG detector followed by a linear SVM), that has been quite popular in recent years, and is the basis of the currently best performing detector. Likewise, for the classification task, a standard and popular bag-of-words approach with a non-linear SVM (Gaussian kernel) was used.

Two objects being “car” and “person” were chosen to facilitate training for the classifier. Each classifier was trained with 500 +ve and 2000 -ve examples for the classification task and 100+ve and 1000-ve for the detection task in relation to each dataset. The tests themselves were performed using 50+ve and 1000 -ve for classification and 10 +ve and 20000 -ve for detection. Each classifier was run 20 times and the results averaged.

Given the manner of training the actual performance numbers are not too meaningful; rather it’s the differences in performance which are telling. Several observations were:

* The best results are typically when training and testing on the same dataset.
* In general, there is a dramatic drop of performance in all tasks and classes when testing on a different test set.

This leads to the conclusion that minimal generalisation is happening when training a system. This is likely the case due to:

* The presence of selection bias.
* There is probably some capture bias.
* There is category or label bias.
* There is the negative set bias.

# **3.2. Negative Set Bias**

A "negative set" in datasets typically refers to a subset of data points that represent examples where a certain condition or attribute is absent or not applicable. This section aimed to determine if the negative sample is representative or sufficient. To facilitate this an experiment to evaluate the relative bias in the negative sets of different datasets was carried out. This aims to find if for instance a “not car” in dataset A is different from “not car” in dataset B.

The idea is to approximate the real-world negative set by a super-set of dataset negatives via combining the negative sets of each of the 6 datasets in the evaluation pool. First, for each dataset, we train a classifier on its own set of positive and negative instances. Then, during testing, the positives come from that dataset, but the negatives come from all datasets combined. The number of negatives is kept the same as the number of negatives of the original test, to keep chance performance at the same level. We ran a detection task with 100 positives and 1000 negatives. For testing, we did multiple runs of 10 positive examples for 20,000 negatives.

For three popular datasets (SUN09, LabelMe and PASCAL) we observe a significant (20%) decrease in performance, suggesting that some of the new negative examples coming from other datasets are confounded with positive examples. On the other hand, ImageNet, Caltech 101 and MSRC do not show a drop. The reasons for this lack of change are likely different for each dataset. ImageNet benefits from a large variability of negative examples and does not seem to be affected by a new external negative set, whereas Caltech and MSRC appear to be just too easy.

# **4. Measuring Dataset’s Value**

There are two ways to improve performance given a particular detection task and benchmark these are:

* Improve the features, the object representation, and the learning algorithm for the detector.
* Enlarge the amount of data available for training.

Although at face value these are simple solutions, they pose a few challenges:

* The first issue is that to achieve a significant improvement in performance, the increase in training data must be very significant (performance has an annoying logarithmic dependency on amount of training data).
* The second issue is that, as discussed in the previous section, if we add training data that does not match the biases of the test data this will result in a less effective classifier.

# **5. Discussion**

When testing a model trained on one dataset using a different dataset, a significant performance drop is not surprising. The decline may result from inadequate object representations and recognition algorithms. These algorithms might overly focus on dataset-specific aspects rather than the core visual task. Human vision copes with biases, but algorithms might struggle. However, algorithms aren't solely to blame. If a dataset defines a "car" as a race-car's rear view, it won't recognize a side view of a family sedan as a "car." Evaluating active recognition datasets, Caltech-101 lacks generalization, MSRC performs poorly, while modern sets like PASCAL VOC, ImageNet, and SUN09 fare better. This indicates progress. Dataset quality matters: for mere feature vector usage, maybe not, but for understanding the visual world, it's crucial. Recommendations for better dataset development follow in the next section.

# **6. Epilogue**

When creating a new dataset to detect and avoid bias, researchers should begin by subjecting the dataset to the tests outlined in their paper. This helps identify main problematic issues early on. To minimize bias during dataset construction:

* 1. **Selection Bias**: Automated dataset collection is better than manual, but internet searches have their biases. Combining data from various sources and countries can help decrease selection bias. Using unannotated images and crowd-sourcing labelling can also help.
  2. **Capture Bias**: Professional and keyword search photos suffer from capture bias, like central object placement or specific orientations. Applying data transformations like image flipping or jittering can reduce this bias. Exploring automatic image cropping is another option.
  3. **Negative Set Bias**: A rich and unbiased negative set is crucial for classifier performance. Datasets focused only on specific objects may lack generalization. One solution is to add negatives from other datasets. Mining hard negatives using standard algorithms and manually filtering them can also be effective, despite potential bias against existing algorithms.

This paper initiates an essential discussion about datasets, acknowledging that biases may still be present. The authors hope to stimulate dialogue about this important and often overlooked issue.