Multimedia Tools and Applications (2019) 78: 9463–9488
https://doi.org/10.1007/s11042-018-6528-x

# Estimating the visual variety of concepts by referring to Web popularity

**Marc A. Kastner[1] · Ichiro Ide[1] · Yasutomo Kawanishi[1] · Takatsugu Hirayama[2] · Daisuke Deguchi[3] · Hiroshi Murase[1]**

Received: 3 February 2018 / Revised: 30 July 2018 / Accepted: 10 August 2018 /
Published online: 23 August 2018


**Abstract**
Increasingly sophisticated methods for data processing demand knowledge on the semantic relationship between language and vision. New fields of research like Explainable AI demand to step away from black-boxed approaches and understanding how the underlying semantics of data sets and AI models work. Advancements in Psycholinguistics suggest, that there is a relationship from language perception to how language production and sentence creation work. In this paper, a method to measure the visual variety of concepts is proposed to quantify the semantic gap between vision and language. For this, an image corpus is recomposed using ImageNet and Web data. Web-based metrics for measuring the popularity of sub-concepts are used as a weighting to ensure that the image composition in a dataset is as natural as possible. Using clustering methods, a score describing the visual variety of each concept is determined. A crowd-sourced survey is conducted to create ground-truth values applicable for this research. The evaluations show that the recomposed image corpus largely improves the measured variety compared to previous datasets. The results are promising and give additional knowledge about the relationship of language and vision.

**Keywords** Visual variety · Language and vision · Concept semantics · Semantic gap

## 1 Introduction

In recent years, the growth of visual data on the Web and in social media is astounding. This results in a need for automated approaches to process such data. Whether the purpose is image retrieval, captioning, or tagging, a comprehensive understanding of image contents becomes crucial. Natural language is vague and the meaning of tagging might change depending on the choice of words. A rather abstract tag like “vehicle” might not describe

Parts of this research were supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research, and a joint research project with NII, Japan.

✉ Marc A. Kastner
kastnerm@murase.is.i.nagoya-u.ac.jp

Extended author information available on the last page of the article.

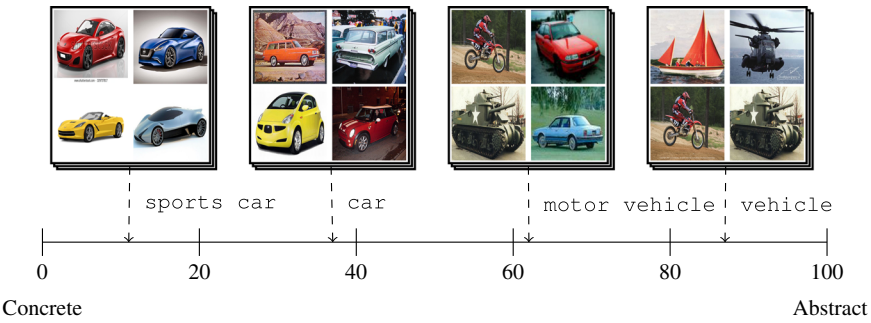
an image of a specific motorbike type particularly well. On the other hand, the model name of said motorbike might be too specific, as an average user might not have a mental image of such a phrase, which would draw the tag pointless. This is a good example that shows that the range of the so-called "semantic gap" lying between language understanding and vision detection could vary. Thus, in order to overcome the semantic gap, it is important to have a deep understanding on how vocabulary and their visual representations connect. Depending on how concrete or abstract a term is, the size of a visual mental image will drastically change. Additionally, the social Web might create a biased view on things. While a very abstract term like "vehicle" might describe all kinds of things —"ships", "airplanes", "trains", etc.— most users would probably only think about "cars" at first glance.

There are existing taxonomies for languages like WordNet [26]. However, these are commonly only based on lexical relations of language like hypernyms and hyponyms. They do not account for the visual features of each concept, and thus it is uncertain whether a lexical taxonomy is equal or even related to its visual properties. For example, imagine two families of animals. The classification of animals is based on biological properties and differences, which are not necessarily related to visual appearance. Thus, one family could have a lot of species which look fairly similar, while another might have very few species with distinct visual differences. The first one would have a larger lexical variety, while the second family would have a higher visual variety.

In this paper, the concept of *visual variety* is introduced as one step to approach the semantic gap. This idea is different from conventional measurements of the distance between visual concepts. This difference is illustrated in Fig. 1. A method to measure the visual variety of language terms is proposed, together with a way to refine the used image corpus to approximate the common mental image for a term or a concept. In this research, one set of images is created for each concept. As shown in Fig. 2a, this method can be used to compute and compare the results for different terms and concepts. The image composition for this is crucial, as it has a large influence on how the visual feature space of the image set will look like. Image sets of abstract concepts are a composition of images of its various sub-concepts. There are sub-concepts closely related to each other and thus often visually very similar, which lowers the score of the overall concept. But there are also sub-concepts which vary visually, and a large number of images of these would increase its score. Thus, how image corpora are composited is crucial for each measurement. For each concept, a


(a) How different are two concepts? (b) In how many ways can it be visualized?

**Fig. 1** **a** shows an example of visual distance. Images of `bicycles` and `motorbikes` are visually similar, but very distinct from images of `cats`. This creates a large distance to `cats`. The visual variety in (**b**) analyzes a single concept. `Sports cars` look visually very similar, as they share similar characteristics. In contrast, a set of images of `vehicles` in general will have a much wider span of visual contents




(a) Comparing the results of visual variety for different concepts and image sets.

(b) Image set composition for the concept `car`. Different colors stand for different subordinate concepts.

**Fig. 2** Core ideas of visual variety. **a** The variety of visual characteristics of a term or concept is quantified. This relates to whether something is concrete or abstract, as an abstract concept has a vague image, often attracting a larger amount of noise and ambiguity. **b** The image set for a single concept, e.g. `car`, is a recomposition of related subordinate concepts. In this image set, the occurrence of images should be roughly equal to a natural distribution of these concepts

well-balanced set of images resembling the common mental image of it is built, as shown in Fig. 2b

To ensure that these image corpora are not biased in an unrealistic way, metrics to determine the popularity of sub-concepts are introduced. Multiple approaches to define popularity are analyzed. For the measurements to yield appropriate results, a distribution which feels reasonable to the majority of people is needed. It is difficult to obtain such a distribution, as it is highly subjective, but a Web-based population distribution is thought to resemble it due to its crowd-sourced nature. Therefore, the core assumption is that the popularity of concepts on the Web approximates the general mental image of these concepts, and thus that there is a direct connection between the visual variety perceived by the majority of humans and Web popularity. In order to approximate a distribution which is related to Web popularity, metrics like analyzing Text or Image Search results are explored. For a comparison, other methods using word frequencies are included in the evaluation. Depending on the metric, one could bias the results, opening opportunities for visual understanding seen from different viewpoints. Lastly, a quantitative analysis compares differently composed image corpus with a crowd-sourced ground truth.

The main contributions of this research are:

– Introducing the concept of *visual variety* as a means of measuring the semantic gap by approximating the common mental image of concepts.
– Proposing a method to compose an image corpus from known datasets using Web population based metrics to create a natural less-biased composition of images.

– Conducting a crowd-sourced survey to obtain ground-truth *visual variety* labels for a selected set of concepts to evaluate the created datasets against how humans perceive each concept.

This paper is structured as follows. Section 2 briefly summarizes existing methods for visual distance measurement. Then Section 3 first describes the method of visual variety measurements using cluster counting. For the approach to yield meaningful results, a well-balanced image corpus is necessary. Therefore, Section 4 proposes a method to construct such an image corpus using Web-based popularity metrics as a weighting. Then Section 5 discusses the crowd-sourced survey used to obtain reasonable ground-truth labels. This is necessary to make a quantitative evaluation of each proposed image corpus. Section 6 shows the evaluation results, which are further discussed in Section 7. The paper is concluded in Section 8.

## 2 Related work

The idea of creating relationships between language and vision dates back a long time. There has been research in computer science, but also in psychology and linguistics. Below, we introduce related research, divided by field or topic.

**Psychology and linguistics** Paivio et al. [30] analyzed the concreteness, imagery, and meaningfulness of nouns. This psychological analysis focused on the differences of literal meanings for different kinds of words. They analyzed whether abstractness and concreteness have an impact on the perception of words. Furthermore, they empirically evaluated the characteristics of a physical representation of a word with its mental representation. Using test subjects, they compared the common mental image of real objects with that of rather theoretical terms.

In the field of linguistics, there have been various approaches for classifying language into a taxonomic structure. The most famous and still common approach for the English language is WordNet [26]. However, this structure is solely based on lexical relationships. It gives a clue on the semantic relationship between concepts and thus is helpful as a basis for this research, but one can not make assumptions on visual features of concepts solely using WordNet.

**Ontology of concepts** Kawakubo et al. [16] proposed an idea how to automatically create an ontology for visual features. They clustered similar images to create a hierarchical structure of related visual concepts. Meanwhile, Inoue et al. [14] tried to analyze the ontologic relationship of visual concepts by directly incorporating the lexical relationships. They calculated a weighting which describes how much a hyponym has a visual influence on its hypernyms.

**Visual concept analyses** Nakamura and Babaguchi [28] measured the distance for visual concepts with an Adaptive Weighting for multiple visual features. Furthermore, Nagasawa et al. [27] analyzed the effect of noise images on distance measurements. They found that in contrast to an image classification algorithm, where any noise often majorly reduces precision, noise images actually have a surprisingly positive effect on distance measurements.

Yanai and Barnard [36] analyzed the image region entropy to calculate the *visualness* of different adjectives. They defined a probability function which decides whether an image region belongs to an adjective or not. Based on this, they compared the amount of entropy for different adjectives and made an assumption how hard it is to visualize them. Kohara and Yanai [19] continued these analyses by looking at adjective-noun pairs.

Divvala et al. [8] proposed a method which uses unsupervised crawling to create a visual knowledge database. By combining related phrases or nouns, a graph of supposedly related visual concepts is created. For every node, an image set was Web-crawled using search engines. To filter-out phrase combinations which have no meaning and result in random images, a classifier is used. Lastly, a clustering is defined using the node-to-node visual distance of neighboring concepts. The automated approach promises to find all concepts related to a starting term.

**Use-cases** Van Leuken et al. [20] did a study on *visual diversification*. The idea is to improve the results of image retrieval by removing similar images of the same object or concept, and thus overall diversifying the retrieved results. In their work, they proposed clustering techniques to create clusters of images which are very closely related. Next, they selected a representative image of each cluster which is used for the image retrieval. As there is no method available to estimate the variety of images, they evaluated their results by comparing the resulting clusters to human-made clusters.

An analysis in the field of psycholinguisitcs by Smolik and Kriz [33] suggests that word imageability and concreteness have a large influence on language comprehension, language production, and language learning. It is thought to be used in both syntactic as well as semantic processes in the human mind. This research suggests, that it is also of high interest for computer-assisted language creation like in natural language processing or image captioning. Li and Nenkova [21] predict sentence specificity using psycholinguistic labels like Imageability, Concreteness, and Meaningfulness. Their research can be helpful to estimate text difficulty or for creating simplified versions of text. However, these metrics are largely based on text semantics and come from a psycholinguistic dataset with only a limited number of generic nouns. Visual variety as a concept could offer extra insight on text semantics from a visual perception point of view. A data mining driven approach to this can furthermore vastly increase the number of labels, making for more precise usage and applications.

Recently, a new field called *Explainable AI* [32] came in closer focus. This field looks at the problem that recent machine learning, especially neural networks, are often black boxes. There is very few insight on how recent advancements work internally, except that they prove to have better accuracy. The field is both interested in how the internals of a trained network work, and in how the results of a classifier are explainable. In the latter, visual variety analyses can find additional insights, which are commonly perceived by a human but yet to be quantified by a machine. For fields where even tiny miscalculations can have a fatal impact, like medicine or aviation, a black box approach can have its issues [12, 13]. Furthermore, in recent advancements of privacy laws, using a black box for machine learning might result in legal issues for business applications. With a similar mindset, in a work by Hentschel and Sack [11], there was an analysis on what data is preserved in Bag of Words classifiers and which image regions are commonly used to detect classes in image classification. These experiments often result in very surprising results, which showcases a mismatch of human perception and computer vision. Such a mismatch is yet to be quantified, but opens the door for additional research on concept semantics and visual features. To this end, visual variety discussed in this paper

quantifies the semantics and differences between different concepts, making results better explainable. As the results are evaluated against human perception, it also closes the semantic gap between humans perceiving their surroundings and the computer analyzing datasets.

In summary, most methods focus on creating a hierarchy or folksonomy of different concepts. While there has been research on concept variety in the field of psychology [30], there is no method which aims to quantitatively estimate the visual variety of general concepts. Having a measurement for visual variety of concepts would provide visual diversification research a ground-truth estimate for quantitative evaluations, as well as help to quantify the semantic gap between human perceived semantics and a machine-trained model.

## 3 Visual variety measurements

Distance measurements are commonly a direct comparison of two visual concepts [28]. The goal is to find the distance between two sets of images and thus trying to make an assumption on how these concepts differ visually. Unfortunately, all those results are relative between the two visual concepts. There is no prediction made on the visual characteristics of a single concept, which creates a gap between language and vision. Related work [1, 36] analyzed the visual entropy of image regions related to adjectives. While they work nicely for adjectives like colors, as they directly describe visual characteristics, there has been less work on how more complex concepts relate to visual variety. In other studies [14], the visual relationship of terms within taxonomies was analyzed. It uses the lexical relationship as a weighting or input value and thus assuming a direct relationship between lexical and visual characteristics. As this is not necessarily true, this assumption would lead to an error when approaching the semantic gap lying between language and vision. There is work regarding visual diversification [20], that aims for a large visual variety in image retrieval result sets. In the evaluation, their approach was compared with a diversification created by humans, in terms of which representative pictures are chosen by each. Unfortunately, the actual effect of this remains unclear, as there were no analysis on how the diversification process influences the dataset and the visual characteristics across it.

Language is naturally created and very complex, which results in word ambiguities and overlaps. A deep language understanding is crucial to solving data analysis problems. In Web and Social Media, visual contents and text are usually co-existing, so the mutual relationship is often used to gain knowledge about data. However, ambiguities make this process prone to mistakes. One can not assume that the visual variety of terms is related to the number of hyponyms or the level of depth within a language taxonomy. WordNet [26] and other taxonomies were not created with any visual aspect in mind, at least explicitly. For example, one family of animals might have a large variety of visual features, colors, size differences, and so on, despite having few species. On the other hand, there might be other families which look closely related to all images, despite having thousands of species, and thus hyponyms. A biological classification and a linguistic taxonomy would have different results than a visual analysis.

In this paper, this gap is approached by an analysis of visual variety. To yield intuitive results, ideally, an image corpus with a comprehensive composition of images which present the common mental image of a concept is needed. In this paper, a set of images is called *balanced*, if there is a meaningful image composition which closely resembles a common variety of a concept. For every concept, an image set with such a balanced composition of images, and its visual vectors, is generated. When looking at the resulting data spatially,

the visual vectors show clusters of very similar concepts. The distance between clusters is the inter-concept distance between visual features, where unrelated images result in a larger distance than closely related images.

When analyzing a very abstract concept like e.g. `vehicle`, a diverse set of images with different kinds of vehicles is intuitively useful. However, a large variety of different cars might have a rather low impact on the mutual distance of image pairs, as these have similar visual features. In contrast, when adding an airplane to the mixture, the distance will be rather high. The *distance* in this case refers to the distance between the visual vectors of each concept's images. Thus, the ratio of how many images of each subordinate concept are within an image set for an abstract term is crucial for the results. For a very abstract concept, like `vehicle`, this creates a variety of spatially distributed clusters in the feature space for sub-concepts like `airplane`, `motor vehicle`, and `ship`. This spatial distribution of visual features is solely based on the visual vector and does not need to be correlated to a lexical taxonomy.

The number of clusters in a spatial clustering relates to the visual variety within a concept. This idea of spatial clustering is visualized in Fig. 3, which shows the visual space of the concept `vehicle` as an example.

For each concept,

$$f(\mathrm{x}) = \#(\mathrm{clusters}(\{\mathrm{features}(i)|i \in \mathrm{images}(\mathrm{x})\})),$$

**Fig. 3** Clustering the visual feature space of the concept `vehicle`. A high visual variety of a concept creates more spatial clusters in the visual feature space. Therefore, clustering algorithms can determine the variety by finding the right number of clusters. Note that this simple approach does not segment or normalize images, so even images of the same vehicle in different situations, applications, or environments can create additional clusters. In contrast, two visually similar images of different sub-concepts might also be clustered together. This behavior is wanted as it is expected to approximate the mental image of the concept more closely

where x is a concept. For a concept x, the visual features in a large number of images are extracted. This visual feature space represents the visual characteristics of the concept, putting similar images spatially closer. The visual features are then spatially clustered, exploiting this idea. The number of clusters are counted, as a high number of clusters indicate non-homogeneous visual characteristics. Furthermore, the more visual characteristics are scattered, the larger the number of clusters get. This equation thus quantifies the spatial scatteredness of the visual feature space, and is comparable between different image sets if the same number of images is used.

## 4 Image corpus construction

Lexical relations within natural languages are commonly described using hierarchical structures. A set of interchangeable synonyms for a specific meaning is commonly called *synset*. The structure of WordNet [26] connects synsets to other synsets by using semantic relations like hypernyms and hyponyms. For example, a rather abstract synset like motor_vehicles might contain more concrete synsets like car and truck which by themselves contain more concrete concepts like sports_car or pickup. As this structure is semantically based on lexical relations, it is uncertain how much it is actually related to visual properties of the underlying visual concepts. ImageNet [7] has a large dataset built on top of WordNet and aims to provide a collection of example images for each concept. It is commonly used as a source of images to train e.g. image classification algorithms. All images were Web-crawled but then filtered by hand using crowd-sourcing techniques. Each synset has between zero and a few thousand images. When emphasizing the hierarchical structure of the dataset, this paper also uses graph theory terminology. In that case, a *root*, *parent*, or *leaf node* refers to a synset, depending on its position in the tree.

### 4.1 Imbalance of WordNet

The experiment starts with a tree extracted from ImageNet. For example, a node called sports_car has a large collection of images of different sports cars. It is a *leaf node*, as there are no hyponyms for this synset in WordNet. This decision is arbitrary and inherited from WordNet. It assumes that different types of sports cars are similar enough, that a further distinction between different models or brands might not be necessary. The rest are *non-leaf nodes* which are usually assumed to be more abstract than leaf nodes. Linguistically speaking, these nodes are hypernyms of the subordinate visual concepts. Non-leaf nodes consist of various visual concepts, described by their hyponyms. An image set for car might contain a number of images of sports cars, albeit not limited to it, as there are also other types of cars. In an even more abstract image set for e.g. vehicles, it might even include tanks, ships, or airplanes. However, do all these sub-concepts have an equal impact on the mental image of a *vehicle*?

The answer is hard to determine, but the assumption is that it relates to how present the individual synsets are in the mental image of its super-concepts. Unfortunately, the crowd-sourced origin of ImageNet often results in a very one-sided set of images. For ImageNet, the goal was to provide an overview of images necessary to grasp its concept. Further analysis shows that leaf nodes can range from very common terms up to rather unknown or obscure terms; e.g. in the *truck* category, there are leaf nodes like moving_van and delivery_truck, which might have a high influence on the common mental image of trucks. In contrast, the same category also contains rather obscure concepts

like `milk_float` (a British milk delivery vehicle) and `book_mobile` (a mobile library), which might not have the same influence on the said mental image.

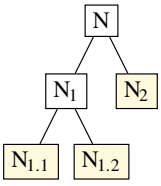
As explained before, the number of hypoynms of a concept can be a misguiding measure for visual variety, as it is a purely linguistic relationship. Similarly, the depth of a term in the tree can be misleading, as narrow concepts like `forklift` are close to `vehicle`, while a similarly concrete `sports_car` has almost double the distance.

### 4.2 Balanced corpus recomposition

The composition of images plays a crucial role for the perceived variety of the image corpus. For this, each non-leaf node image set is recomposited based on images of its hyponyms. Starting from a given root node, a full WordNet sub-hierarchy is extracted. Next, a list of representing synonyms for each synset is accumulated. This can vary from different spellings (British vs. American) up to other words which are interchangeable but commonly have the same meaning when used in a related context (e.g. *cab* and *taxi*).

To make this recomposition well-balanced, a distribution function defines the ratio of images used from each hyponym. The distribution function aims to select an image composition which feels natural for the majority of people. Therefore, it looks at how popular a term is within its group of related concepts, to determine how relevant a sub-term is in the mental image of this concept. As a metric for term popularity, there are a couple of options. The API from common search engines may serve as a Web-based approach to measure the popularity of terms. Using the Google API, it is possible to crawl an approximation of total results for either text or image search results. It is also possible to use a metric based on word frequencies. This is a common approach used in linguistics to compare the popularity of different terms, adjusted for grammatical suffices. Using this, large amounts of text can be searched for the number of occurrences, a term or phrase appears.

Applying such a metric, a *popularity score* for each synonym for each synset is chosen. In Section 7.1, a variety of metrics are compared more extensively. As multiple synonyms of the same synset usually have a large overlap, the average of its popularity scores is used to describe the popularity of this synset. The non-leaf image sets are merged together using the previously determined ratio, as explained in Fig. 4. This is believed to be superior to




(a) Getting hyponym leaf nodes of synset $N$.

| | |
|---|---|
| Popularity of $N_{1.1} \in N$ : | $w_1$ |
| Popularity of $N_{1.2} \in N$ : | $w_2$ |
| Popularity of $N_2 \in N$ : | $w_3$ |

(b) Determine weighting using popularity metric.

$$N' = w_1 p(N_{1.1}) + w_2 p(N_{1.2}) + w_3 p(N_2)$$

(c) Re-composing the image corpus. $p$ is the function for retrieval of synset images.

**Fig. 4** Recomposition of the imageset for a synset $N$. First, in (**a**) all relevant hyponym leaf nodes for a synset $N$ are extracted from WordNet. Then, in (**b**) a weighting for each hyponym relative to its parent is determined using a popularity metric. Lastly, a new imageset $N'$ for the synset $N$ is recomposed with an appropriate number of images from each hyponym, as shown in (**c**). This procedure is repeated for every non-leaf node in the WordNet tree

a crawling of non-leaf node images, as the composition of images would be uncertain and hard to validate.

### 4.3 Increasing the amount of images

The number of images available in ImageNet vastly varies depending on the synset. There are synsets which have rather obscure terms, so it is hard to find fitting images for these visual concepts. For these synsets, ImageNet provides either none or a minuscule amount of images. Assuming that these terms are either too vague or too obscure to have an influence on more abstract image sets, they are removed from the hierarchy.

As the non-leaf node image sets are composited from multiple leaf nodes, the amount of leaf node images becomes a major bottleneck. Extra images are crawled using Search Engine API [10, 25] to increase the number of images. By combining synonyms for each synset, the number of crawlable images can be increased. To make the results more relevant and decrease the major reason for noise, a common phrase describing all synsets can be appended. For example, when crawling images related to `car`, `truck`, and `motorbike`, appending *vehicle* to each search might be a simple approach to decrease a certain amount of completely unrelated images. The full process of image corpus construction is visualized in Fig. 5.

Of course, Web-crawled approaches introduce a very high ratio of noise. Kennedy et al. [17] suggest a more than 50 percent chance of noise, even for dedicated image services like Flickr [35]. For Google Image Search [10], the ratio seems to be even worse, but highly depending on the search term. However, the noise is not necessarily a negative thing. While it is intuitive that noise images have a negative impact on image recognition algorithms, this conclusion might not hold true for visual variety measurements [27]. We consider that there is a semantic relationship which corresponds to why the noise exists in the first place and thus removing noise images could also remove hidden semantics. Therefore, there is no further attempt to filter out noise images in our work.

## 5 Obtaining the ground truth

The goal of this research is the measurement of visual variety in a common mental image. Each term would have a value attached which describes its average visual variety, on where

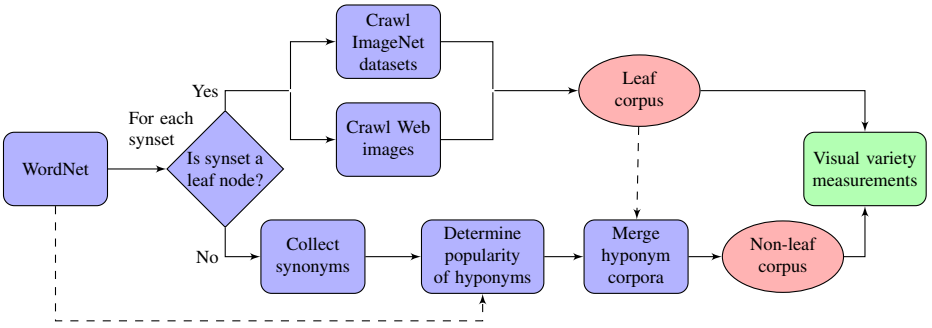



**Fig. 5** Image corpus construction. The leaf image sets are constructed by using a combination of ImageNet and Web-crawled images. Any non-leaf node is composited from leaf-nodes incorporating the popularity of each hyponym. This ensures the resulting corpus to be close to the common mental image. This is done for each node individually

a majority of people would agree. While this is rather subjective, it is expected to achieve stable results in a majority decision when including a sufficiently large number of people. To the best of our knowledge, there exists no dataset with this kind of labeling. Therefore, to make a quantitative analysis of the proposed method possible, an excerpt of WordNet has been annotated with visual variety labels.

### 5.1 Crowd-sourced survey

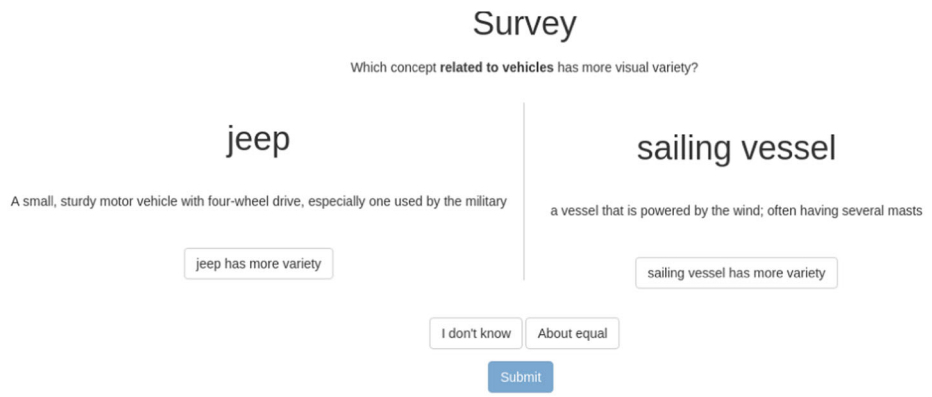
To form a reliable ground truth for this rather subjective measurement, a large enough number of people needs to be asked. Therefore, a crowd-sourced survey using Thurstones' method of pair comparisons [34] has been conducted. In Thurstones' method, survey participants are shown only two samples of a larger set of objects at a time. They are asked to answer a question comparing these objects.

Thurstones' method is in particular useful for hard-to-decide questions of individual preferences. Assuming the ordering is transitive, a ranking can be obtained after asking the participant about a sufficient number of pairs. This exploits the fact that it is often easier to choose between two than choosing between many.

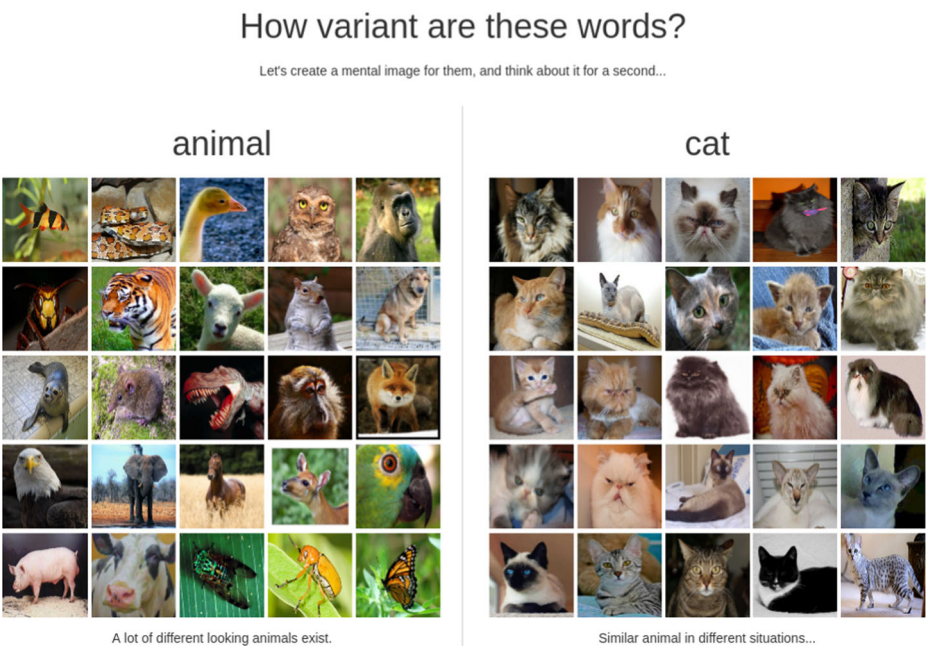
As this method is ideal for subjective questions which are hard to decide, it adapts well to visual variety. We set up a survey to conduct such an experiment for this research. On each page, a participant sees the name (e.g. "`vehicle`") and a short description (e.g. "a conveyance that transports people or objects") of two synsets. They are asked to visualize these concepts in their head and decide, which one is more visually variant. The participants are asked to make that judgement without further researching either concept, but by just making an assumption based on their prior knowledge on them. Note that no visuals or images are shown to avoid biasing the results in a predefined direction. To avoid confusion and misjudgement based on knowledge, all chosen synsets are commonly known terms.

For every pair comparison of concepts `A` and `B`, a participant can choose one of four buttons: "`A` has more variety", "`B` has more variety", "About equal", or "I don't know". The user interface is shown in Fig. 6a. The first two buttons are asked to be pressed when a participant considers that either of the two concepts has a larger visual variety. The "About equal" button is for the case where a participant cannot make out which one is even slightly more variant. Last, the "I don't know" button is a skip button for the case where a participant does not know either or both of the concepts and thus is unable to make a judgment. In the introductory text, it is emphasized that either of the latter two buttons should be used as little as possible. This is to avoid over-selecting the "About equal" button, as quite a few comparisons can be difficult for most participants.

The concept of visual variety is novel and thus hard to convey. Therefore, the introduction of the survey starts with a short tutorial. In this tutorial, the concept of visual variety is explained by showing examples. These examples use a different set of synsets, which are not part of the main survey. First, the tutorial shows a pair comparison, just as the main survey would. After selecting either button, the participant proceeds to a page, where a variety of pictures for both synsets are shown. This is to show participants what they are supposed to visualize in their minds. The pictures were handpicked with the goal to make it clear what visual variety is supposed to mean. Figure 6b shows an example of the tutorial page explaining the synsets `animal` and `cat`. Afterwards, the tutorial goes back to showing the participant the previous pair comparison, outlining which button would be the recommended solution for this pair (e.g. `animals` have more visual variety than `cats`.) All examples in the tutorial are chosen to be rather extreme, so most partcipants would likely agree with




(a) Survey: Main part




(b) Survey: Tutorial

**Fig. 6** **a** shows the user interface of the main survey. The participants are asked to make a judgement on the visual variety of a pair of two concepts. Before the main survey is taken, a tutorial as shown in (**b**) will explain the idea of visual variety and what a participant is supposed to visualize in his/her mind

these recommendations. The tutorial shows four such example pairs, each with a selection of pictures to outline the way of visualizing them in ones head. They include an example of an "about equal" edge case, as well as an example of a surprising outcome. After the tutorial is finished, the main survey proceeds as explained before.

### 5.2 Results

Over the course of two months, the survey has been promoted through Web and Social Media including Facebook, Twitter, and Reddit. Compared to solutions like Amazon Mechanical Turk (AMT) [3], this has the effect that mostly volunteers are participating surveys. As participants are not paid, this can decrease the risk of spammers and thus improving the quality of results. Largely, a majority of replies seemed to take the survey diligently — most results match and people took a reasonable amount of time for answering each pair comparison. There were, however, a small number (around 5%) of dubious cases where people replied the survey suspiciously. Here, people evidently skipped most explanations and the time taken per pair comparison became significant outliers compared to others. As these responses also usually did not match the responses of other participants, suspicious results were treated as spam and filtered-out.

The survey was carried out in English and publicly available in crowd-sourced manner. While there was no restriction to native speakers, we asked participants to only participate if they are confident enough in their English proficiency. For the main survey, 25 synsets related to vehicles have been chosen. They span a variety of levels of abstractness (such as `vehicle`, `motor vehicle`, `car`, and `sports car`) and areas (such as `street vehicles`, `air vehicles`, `water vehicles`, and `war vehicles`). Each synset was labelled with a valid description fitting the WordNet meaning of the concept. The descriptions were sourced from Merriam-Webster's Dictionary [24], Oxford Dictionary [29], and WordNet itself. They were selected to have a similar detail and length for each synset to reduce visual bias on the survey pages themselves.

After finishing the tutorial, each participant was asked to judge thirty pair comparisons. Voluntarily, participants were able to extend the survey, in which case more unique pair comparisons would have been shown, but only one participant chose to do so. Likewise, any participant was able to stop the survey at any point, in which case only the pair comparisons up to that point have been saved to the database.

In total, 158 people participated, answering 4,529 pair comparisons (avg. 28.66 per participant and 13.36 answers per pair.) Out of these, 442 answers were pairs considered equally variant and 63 comparisons were skipped with the "I don't know" button. Each pair comparison in average took 8.35 seconds. Out of all pairs, 87% reached a majority for either one of the two concepts. There were two pairs, where one of the skip buttons gained a majority.

In the 13% of problematic pairs without a majority, there were a couple of noticable patterns. First, there were pairs, where both concepts were rather concrete leaf nodes in different sub trees. `bicycle` vs. `motorcycle` is already pretty hard to decide, but when comparing either to a `warship`, people might just give up and click something randomly. Therefore, it is actually surprising, that the greater number of pairs could reach a common majority.

On a similar note, there are synsets which are hard to understand, or may even be misconceptions. One particularly ambigious synset is `self-propelled vehicle`. The synset basically contains vehicles using a motor, but is different from the synset `motor vehicle`, which only contains *road* vehicles using a motor. This semantic nuance is inherited from WordNet and unknown by most participants. Therefore, it can lead to a confusion and nonhomogeneous results.

## 6 Experiment

The aim of this research is to measure the visual variety of concepts as judged by the majority of people. Previously, a naive method to measure visual variety using cluster counting has been discussed. Due to its simplicity, it depends on a well-balanced image corpus to lead to meaningful results. Therefore, we proposed a method to create such a corpus using popularity metrics. To evaluate these methods, they are compared to ground-truth values obtained from the conducted crowd-sourced survey.

### 6.1 Image corpus creation

For the evaluation, a plain ImageNet serves as a baseline. This will outline, how well (or rather, badly) an unmodified downloaded copy of ImageNet performs in visual variety measurements using the cluster counting method. The other three methods are modified and recomposed versions of the plain ImageNet.

Based on WordNet, a tree of about 600 nodes starting from the root node `vehicles` was extracted using NLTK [22]. Leaf nodes with a very tiny amount of images were removed, with the remaining tree resulting in about 800 to 1,500 images per node. The aim is for an equal amount of images in every node.

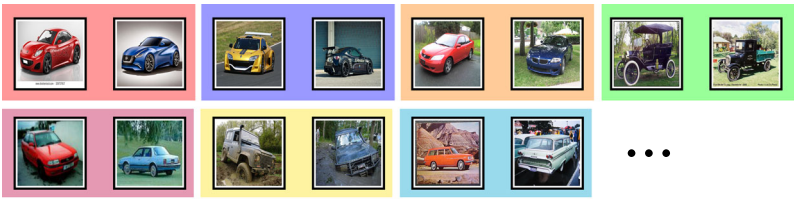
As the ground-truth results of the survey span 25 core synsets, the goal was to obtain a decent amount of images for each of them. Note that there is a larger number of nodes still influencing the composition of each parent synset's image corpus, even if not chosen for direct evaluation. To increase the available visual data, Google [10] and Bing [25] APIs were used for additional crawling of Web images. Potential duplicates are deleted by image comparison.

For evaluating the proposed method, the image corpus of all non-leaf nodes is recomposed using two Web popularity metrics, and some extra images are added using Web-crawling. The Google API [10] has been used as a metric to approximate the Web popularity of various sub-concepts. For each term, the maximum amount of search results for either the Google Text Search (Proposed method 1) or Google Image Search (Proposed method 2) serve as metrics for the recomposition. These numbers reflect the common popularity of terms within the indexed Web content. A discussion in Section 7.1 will go into greater detail on how different metrics for Web Popularity affect the recomposition of the image corpus.

Lastly, as a comparative method, an equally weighted corpus has been created. Here, all leaf nodes influence a parent node equally. This means, the structure of WordNet is inherited and a parent node receives the same amount of images from each of its leaf nodes.

An example of the resulting image corpora for the synset `car` is visualized in Fig. 7. In the top, an equal weighted distribution (Comparative method) is used to produce an image composition where each subordinate concept is treated equally. In contrast, the bottom rows show compositions where the Google Text (Proposed method 1) and Google Image (Proposed method 2) popularity metric create more natural distributions.

Due to the different ratios for each composition method, it was not possible to reach the same volume of images per synset for each corpus. A higher number is favorable, so the highest common volume of images in all synsets per corpus was chosen for further evaluations. Accordingly, the Baseline dataset uses 1,000 images per synset, the corpus for Proposed method 1 has 2,000 images, and both the Comparative corpus and the Proposed method 2 corpus contain 2,430 images per synset each. As the Plain ImageNet dataset is an unmodified copy of the original ImageNet, there were no means taken to increase its amount of data.



(a) Composition of synset `car` corpus using equal weighting (Comparative method.)

(b) Composition of synset `car` corpus using the Google Text-based distribution (Proposed method 1.)

(c) Composition of synset `car` corpus using the Google Image-based distribution (Proposed method 2.)

**Fig. 7** Parent node composition. Each colored block represents a different subordinate concept merged into the parent node image corpus. **a** shows the resulting composition for `car` when all leaf nodes are treated equally. This corpus is used as a comparative method. **b** and (**c**) show compositions for `car`, where the distribution is based on the Google Text (Prop. method 1) or Google Image (Prop. method 2) popularity metric. The raw values for these metrics are shown in Table 3b

### 6.2 Survey results

Based on the results from the survey discussed in Section 5, the ground truth has been obtained. Each answer by a participant is added to a weighted directional graph, where each node is a synset and an edge describes the difference of variety between two nodes. Answers where "I don't know" or "About equal" were chosen, are skipped.

The resulting graph is put into a maximum likelihood estimation to determine a ranking using Choix 0.3.0 [23]. For further steps, the ranking has been normalized between 0 and 100, where 0 would be the most concrete concept and 100 the most abstract one. The ranking for the ground truth is listed in Table 1.

### 6.3 Measurement results

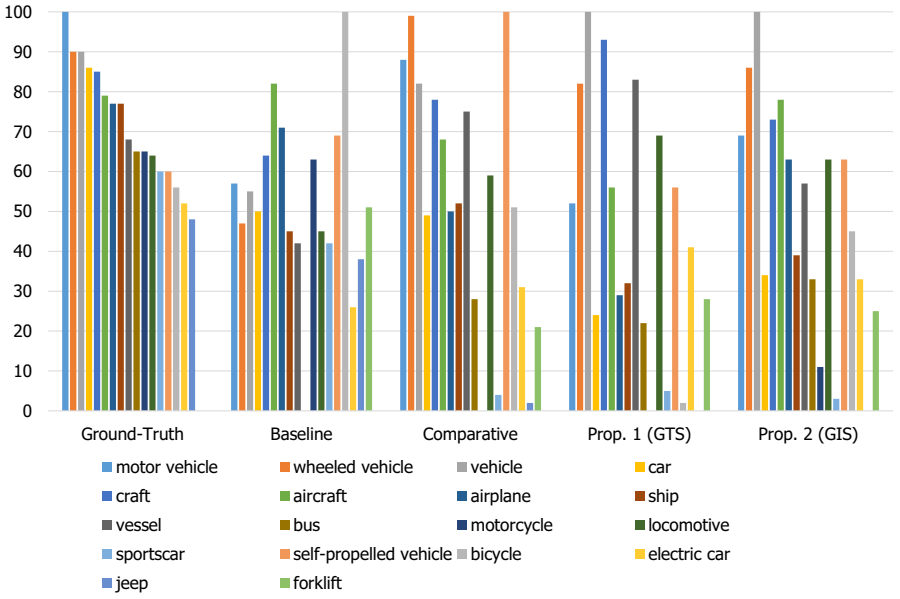
The evaluation examines the data clustering of each image set, as previously discussed in Section 3. For each synset, the number of clusters within the visual feature space of the synset's images represents the visual variety.

**Table 1** The variety results for all compared methods

(a) Visual variety results for each synset, normalized between 0 and 100

| Synset | GT | Baseline | Comparative | Prop. 1 (GTS) | Prop. 2 (GIS) |
|---|---|---|---|---|---|
| motor vehicle | 100 | 57 | 88 | 52 | 69 |
| wheeled vehicle | 90 | 47 | 99 | 82 | 86 |
| vehicle | 90 | 55 | 82 | 100 | 100 |
| car | 86 | 50 | 49 | 24 | 34 |
| craft | 85 | 64 | 78 | 93 | 73 |
| aircraft | 79 | 82 | 68 | 56 | 78 |
| airplane | 77 | 71 | 50 | 29 | 63 |
| ship | 77 | 45 | 52 | 32 | 39 |
| vessel | 68 | 42 | 75 | 83 | 57 |
| bus | 65 | 0 | 28 | 22 | 33 |
| motorcycle | 65 | 63 | 0 | 0 | 11 |
| locomotive | 64 | 45 | 59 | 69 | 63 |
| sports car | 60 | 34 | 4 | 5 | 3 |
| self-propelled veh. | 60 | 69 | 100 | 56 | 63 |
| bicycle | 56 | 100 | 51 | 2 | 45 |
| electric car | 52 | 26 | 31 | 41 | 33 |
| jeep | 48 | 38 | 2 | 0 | 0 |
| forklift | 0 | 51 | 21 | 28 | 25 |

(b) A plot of the results in Table (a), visualizing the overall trend of each dataset




The results reflect the visual variety of each synset. A high number indicates a synset with which is rather abstract, which has a high variety of visual characteristics spanning different images. A low number is a more concrete synset, which has rather homogeneous visual characteristics. The ground truth (GT) refers to the results from the crowd-sourced survey. Baseline refers to the plain ImageNet, while Comparative and Proposed 1/2 refer to our image corpus recomposition method with different weightings. For Comporative, all hyponyms of a synset are weighted equally, while Proposed 1 and 2 use a weighting based on Google Text Search (GTS) and Google Image Search (GIS) results, respectively. The table is sorted in order of the ground-truth ranking

The implementation uses OpenCV 3.2 [15] for feature extraction and distance measurements, and Scikit-learn 0.19.0 for clustering [31]. For each image, the visual features are extracted in form of a Bag of Words model using SURF descriptors [2, 5]. A mean-shift clustering [4] is used to create a clustering of the visual vectors. Then, the number of clusters for every synset is counted. Lastly, they are normalized between 0 and 100 to allow a rank comparison to the ground truth. This process is repeated for all four corpora created. Table 1 shows the ranking results for each corpus.

### 6.4 Rank comparison

A comparison of the ranking generated for each corpus with the ground truth can be seen in Table 2. As metrics for evaluation, the Spearman Rank Correlation [9] and the Mean Squared Error (MSE) have been chosen.

The proposed method 2 using "Google Image weighting" is leading the Rank Correlation with an improvement of 17.7% over the comparative method "Equal weighting" and 192% over the baseline.

The Baseline using the Plain ImageNet has a very low rank correlation. This suggests that the results are scattered and do not fit the crowd-sourced results. When comparing the rankings in Table 1, one can see that the Baseline ranking for each synset is very similar. As a matter of fact, if skipping the normalization, the raw amount of clusters found for each image set is almost identical, so all rankings gather around a similar, rather random, value. Thus, there is almost no correlation, but a surprisingly low MSE, as the average error is relatively low.

The Comparative method "Equal weighting" is a strong improvement over the baseline, although it can not reach the accuracy of the proposed method 2. It uses no weighting, but inherits the distribution from the structure provided by WordNet. The prominent change shows, how crucial the image corpus composition is for the visual variety measurement. Unfortunately, the Google Text weighting worsened the results, as it highly increased the MSE.

For both the Proposed methods and the Comparative method, the MSE seems to be a smaller improvement than the Rank Correlation. They result in a more diverse ranking, and thus, wrongly classified results will have a larger impact on the MSE.

**Table 2** Quantitative analysis of the measurements against ground truth

| Corpus | Rank correlation (larger = better) | Mean squared error (lower = better) |
|---|---|---|
| Plain ImageNet (Baseline) | 0.25 | 10.54 |
| Equal weighting (Comparative) | 0.62 | 9.23 |
| Google Text weighting (Proposed 1) | 0.56 | 14.89 |
| Google Image weighting (Proposed 2) | **0.73** | **9.01** |

This table shows the Spearman rank correlation and Mean Squared Error between each evaluated corpus and the ground truth based on the crowd-sourced survey. As shown, the Proposed methods have a strong lead in either category. The Plain ImageNet dataset has a low correlation as it results in an almost identical amount of clusters for almost every synset

## 7 Discussion

The previous evaluations in Section 6 looked at how recomposed image corpora compare to a conventional dataset for visual variety measurements. It shows, that a recomposition has great potential for improving the measurement. The following will first analyze how the choice of different popularity metrics can influence the results. The Google API metrics used in the evaluation are compared with two alternative candidates. Lastly, other difficulties of the recomposition and obtaining a viable ground truth are discussed.

### 7.1 Different popularity metrics

The proposed method heavily relies on the used image corpus as its composition is crucial for the algorithm to yield meaningful results. The following will discuss four different metrics for popularity. Using one of these metrics, the corpora can be recomposed using the ratio of how popular its leaf nodes are relative to each other.

The first two metrics use the Google API [10], where the maximum amount of search results per term is used as a metric for how popular terms are relative to each other. This reflects the common popularity of terms within indexed Web content. Thus, it makes an assumption on the expectation of image contents in social media. The API provides data for both text and image searches, so they are evaluated separately. These metrics were used in the previous experiment in Section 6.

Third, the Sketch Engine (SE) [18] provides a large Web-crawled text corpus consisting of 19 billion words. This is another fully Web data-based approach, from a different viewpoint than Google results. It is not directly affected by SEO keywords (Search Engine Optimization) or Google PageRank, and solely relates on crawled text-only data. Lastly, the Corpus of Contemporary American English (COCA) [6] provides a large English text corpus with currently 520 million words. It is said to be a well-balanced combination of written texts from newspapers, journals, magazines, and transcripts. Thus, this metric is a non-Web data based comparison.

In the following, we will compare the ratio found by each of these four methods. Table 3a shows the distributions for the synset `truck`, while Table 3b those for the synset `car`. For the synset `car`, the Web-based approaches often composite results in a strong bias towards `sports car`. There is a vast amount of sports car images on the Web for marketing purposes and social media, and thus `sports car` is a category where people intuitively are more likely to upload images to the Web. Therefore, the expectation of an image of a `car` might actually have a strong bias towards `sports car`. The sub-tree related to `truck` is more balanced towards multiple hyponyms. Overall, the Google Search results, especially the Image Search results seem to be the best fit for the visual variety measurements, as they fit the expectations the closest.

### 7.2 Difficulties in corpus construction

Unfortunately, seven synsets selected for the crowd-sourced survey turned out to be hard to crawl. This includes a number of synsets from the non-ground vehicle subtree of `vehicle`, for example `sailing vessel`, `cargoship`, `warship`, and `warplane`. Even after including extra data from other search engines, they resulted in a substantially fewer number of images than the rest of the synsets. Therefore, they were skipped in the evaluations.

Depending on the chosen Web popularity metric, a single leaf node can become an outlier in popularity. This can be seen in the previous example of the synset `sports car`,

**Table 3** Different Web popularity measurements

| Leaf node | GTS | GIS | SE | COCA |
|---|---|---|---|---|
| (a) Distribution of the synset `truck` | | | | |
| `moving_van` | **22.8%** | **27.4%** | 2.4% | 1.4% |
| `delivery_tr` | 9.6% | **23.7%** | 1.8% | 0.9% |
| `pickup` | **14.7%** | **10.9%** | 1.7% | **44.0%** |
| `trailer_tr` | 7.1% | 8.5% | 2.5% | 5.8% |
| `fire_engine` | **11.4%** | 6.8% | 1.0% | 2.6% |
| `tractor` | 6.8% | 6.0% | **12.8%** | **26.8%** |
| `police_van` | 9.8% | 4.2% | **58.4%** | **10.7%** |
| `milk_float` | 1.8% | 2.6% | 0.3% | 0.0% |
| `transporter` | 2.6% | 2.1% | 0.6% | 1.6% |
| `lorry` | 1.9% | 2.2% | **7.8%** | 1.0% |
| (b) Distribution of the synset `car` | | | | |
| `sports_car` | **32.5%** | **27.4%** | **45.7%** | 1.2% |
| `racer` | 6.7% | **9.2%** | 0.3% | 2.3% |
| `model_t` | **24.0%** | **8.8%** | 0.8% | 1.3% |
| `coupe` | 2.3% | 6.9% | 3.5% | 3.6% |
| `used-car` | **11.0%** | 6.7% | 0.4% | 1.8% |
| `jeep` | 1.8% | 5.0% | 1.3% | 6.4% |
| `beach_wagon` | 2.2% | 4.8% | 2.5% | **6.7%** |
| `compact` | 3.3% | 4.5% | 0.4% | **11.0%** |
| `cab` | 1.9% | 3.9% | **3.4%** | **13.3%** |
| `hatchback` | 2.7% | 1.2% | **11.4%** | 1.1% |
| `ambulance` | 1.4% | 0.6% | 0.8% | **15.9%** |
| `minivan` | 1.3% | 0.7% | **8.5%** | 4.8% |

This is an example of how Web popularity distributions will affect the recomposition of parent nodes from leaf nodes for (a) `truck` and (b) `car`. Google Text Search (GTS) / Google Image Search (GIS) are Web-based and show a bias towards social media. The values for both Sketch Engine (SE) and the COCA text corpus are given as a comparison. The bold values refer to the top three of each method

which becomes 45.7% of `car` images for the Sketch Engine (SE) metric (Table 3b). In such extreme cases, the amount of available leaf node images often bottlenecks the retrievable images for parent node corpora, even up to a much higher level in the hierarchy like `vehicle`.

On a similar note, many nodes of ImageNet initially have none or very few images. They can be excluded to simplify the recomposition, but this inevitably results in less variety for the recomposition of parent nodes and thus some introduced bias.

### 7.3 Ground truth results

When looking into the raw results of the ground truth, it becomes evident that there is a bias for objects which are more present in daily life. For example, the synset `car` is one of the

highest ranked synsets, despite pragmatically thinking being rather concrete compared to many other concepts.

To see whether the number of participants is sufficient, the stability of results in relation to the number of participants has been investigated. For this, the resulting rank correlation for different numbers of participants has been sampled between one and 150 participants. Each datapoint represents the average of 15 samples over all participants. The results are shown in Fig. 8. As seen, a tendency of the final results are determined rather quickly. With more participants and the results getting more refined, the results for the proposed methods gain a stronger lead.

### 7.4 Applications

As other work suggests [21], measurements of imageability and concreteness can be used to estimate text specificity. This is useful, if judging ease of reading, or for text simplification. Unfortunately, existing datasets of similar metrics are flat and rather small. A comparison of related terms like `car`, `sportscar`, `motor vehicle`, and `vehicle`, would be difficult, as labeling for hyponym or hypernym concepts is usually missing. Thus, this research could be used to vastly increase the training data samples for such kinds of applications.

In the field of *Explainable AI* [32], where the goal is to bring light into black-boxed approaches in machine learning and artificial intelligence, the results could also be of interest. Over all, this approach quantifies the semantic gap, which shows hidden semantics based on human perception. Thus, it can be a clue on *why* image classification acts the way it does. In a work by Hentschel and Sack [11], it was found that the contents of

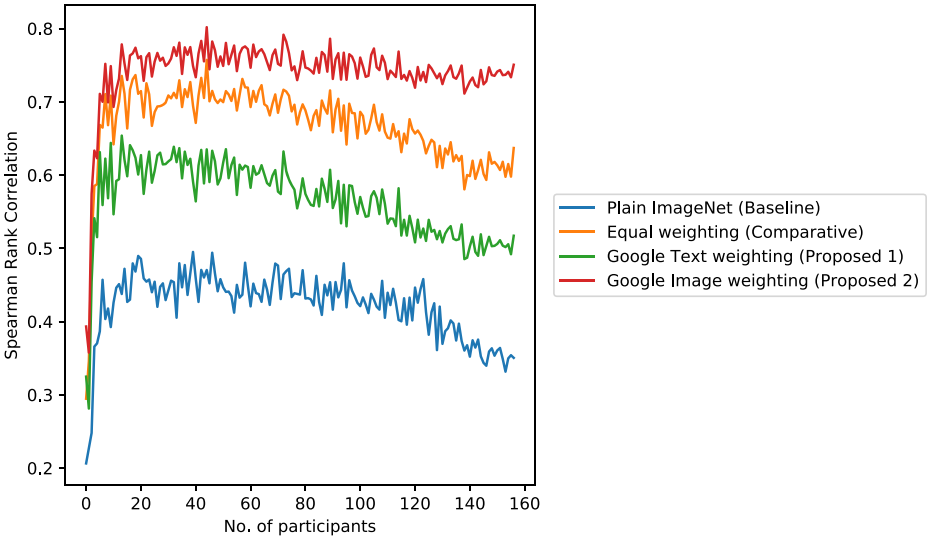



**Fig. 8** The stability of Spearman Rank correlation results for different ground-truths. This figure shows how the Spearman Rank correlation changes with the number of participants used for calculating the ground-truth. For each point, it shows an averaged sample of 15 samples as base for calculating the ground-truth values. It shows that when including about 30 participants, the general trend starts to develop. The gap between each approach and the Baseline can be seen, but the difference between the Comparative and Proposed methods (especially Prop. 2) are lost in noise. After reaching more than 100 participants, the final results can already be approximated closely

image classification models are often very surprising when inspected from a human viewpoint. Trained models find something different than the human would expect them to find, despite often having a very high precision. This can often lead to unexpected behavior for new images, but also showcases the semantic gap between human perception and computer vision. Therefore, our research is thought as an assistance to these and related semantic problems.

## 8 Conclusion

In this paper, a method to measure the visual variety of terms has been proposed. Web data is used to create and enhance an image set for each term based on popularity in social media. The cluster counting method calculates a distinct value for every term, describing its visual variety. Using a crowd-sourced survey, a ground truth for this purpose has been obtained. When comparing the proposed image corpora with another, it shows that the correlation to ground truth highly depends on the used recomposition. The results are promising in terms of understanding the relationship between language and vision.

The presented work approaches the semantic gap by rating the perceived variety of concepts. Therefore, this is valuable as training data for image captioning and tagging approaches. By including data on how words are perceived by users, captioning results can achieve a more natural usage of language. As the semantic gap varies for different cultures and languages, the weighting used for recomposition can bias the results for such purposes. With an increased knowledge of the language, these results could therefore prove beneficial for machine translation and natural language processing. This thought is strengthened by recent studies in psycholinguistics [33] and the use of similar metrics in language models [33]. A better text understanding thus can improve language models further. There has also been advancements in the field of explainable AI, looking into understanding and visualizing artificial intelligence models [11, 32]. In a similar mind set, our method can find hidden data set semantics, especially when considering human perception.

The perception of abstractness for certain words could be compared to choose similarly concrete words for translations or descriptions. The metric used for recomposing the image corpus turns out to be a parameter which is able to influence the overall bias of an image corpus. One could compare the visual image of concepts between different professions or from different cultures and thus use it for market analyses purposes.

For future work, including the metadata of crawled images can provide useful additional information. A combination of the number of clusters and the distances of images inside them can be used to further enhance the results. Other approaches like region segmentation can be used to weight the influence of background and foreground contents.

As shown before, the weighting has a high influence on the results of the variety measurements. Therefore, it could be used to bias the corpus in different direction; for example, as seen from different political viewpoints, communities, or professions. Another interesting comparison would be cultural influence on the results, as well as whether native speakers have a different image of terms than non-native speakers.

**Acknowledgements** We are grateful to Dr. Kazuaki Nakamura at Osaka University who provided expertise that greatly assisted this research.

## References


1. Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM, Jordan MI (2003) Matching words and pictures. J Mach Learn Res 3:1107–1135. https://doi.org/10.1162/153244303322533214
2. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-Up Robust Features (SURF). Comput Vis Image Underst 110(3):346–359. https://doi.org/10.1016/j.cviu.2007.09.014
3. Buhrmester M, Kwang T, Gosling SD (2011) Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspect Psychol Sci 6(1):3–5. https://doi.org/10.1177/1745691610393980
4. Comaniciu D, Meer P (2002) Mean Shift: A robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24(5):603–619. https://doi.org/10.1109/34.1000236
5. Csurka G, Dance CR, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Proceedings of ECCV 2004 Workshop on Statistical Learning in Computer Vision, pp 1–22
6. Davies M (2008) The corpus of contemporary American English: 520 million words, 1990–present. http://corpus.byu.edu/coca/
7. Deng JDJ, Dong WDW, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: Proceedings of 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 2–9. https://doi.org/10.1109/CVPR.2009.5206848
8. Divvala SK, Farhadi A, Guestrin C (2014) Learning everything about anything: Webly-supervised visual concept learning. In: Proceedings 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp 3270–3277. https://doi.org/10.1109/CVPR.2014.412
9. Dodge Y (2008) Spearman rank correlation coefficient. In: The Concise Encyclopedia of Statistics. Springer, New York, pp 502–505,. https://doi.org/10.1007/978-0-387-32833-1_379
10. Google (2016) Google Custom Search API. https://developers.google.com/custom-search/
11. Hentschel C, Sack H (2015) What image classifiers really see —visualizing bag-of-visual words models. In: Advances in Multimedia Modeling: 21st International Conference on Multimedia Modeling Processing. Springer, Lecture Notes in Computer Science, vol 8935, pp 95–104. https://doi.org/10.1007/978-3-319-14445-0_9
12. Holzinger A, Biemann C, Pattichis CS, Kell DB (2017) What do we need to build explainable AI systems for the medical domain?. Computer Research Repository arXiv:1712.09923
13. Holzinger A, Malle B, Kieseberg P, Roth PM, Müller H, Reihs R, Zatloukal K (2017) Towards the augmented pathologist: challenges of explainable-AI in digital pathology. Computer Research Repository arXiv:1712.06657
14. Inoue N, Shinoda K (2016) Adaptation of word vectors using tree structure for visual semantics. In: Proceedings of 24th ACM Multimedia Conference, pp 277–281. https://doi.org/10.1145/2964284.2967226
15. Itseez (2015) Open source computer vision library. https://opencv.org/
16. Kawakubo H, Akima Y, Yanai K (2010) Automatic construction of a folksonomy-based visual ontology. In: Proceedings of 2010 IEEE International Symposium on Multimedia, pp 330–335. https://doi.org/10.1109/ISM.2010.57
17. Kennedy LS, Chang SF, Kozintsev IV (2006) To search or to label?: Predicting the performance of search-based automatic image classifiers. In: Proceedings of 8th ACM International Workshop on Multimedia Information Retrieval, pp 249–258. https://doi.org/10.1145/1178677.1178712
18. Kilgarriff A, Baisa V, Bušta J, Jakubíček M, Kovávr V, Michelfeit J, Rychlý P, Suchomel V (2014) The sketch engine: Ten years on. Lexicography 1(1):7–36. https://doi.org/10.1007/s40607-014-0009-9
19. Kohara Y, Yanai K (2013) Visual analysis of tag co-occurrence on nouns and adjectives. In: Li S, El Saddik A, Wang M, Mei T, Sebe N, Yan S, Hong R, Gurrin C (eds) Advances in Multimedia Modeling: 19th International Conference on Multimedia Modeling Processing, vol 7732. Springer, Lecture Notes in Computer Science, pp 47–57. https://doi.org/10.1007/978-3-642-35725-1-5
20. van Leuken RH, Garcia L, Olivares X, van Zwol R (2009) Visual diversification of image search results. In: Proceedings of 18th International Conference on World Wide Web, pp 341–350. https://doi.org/10.1145/1526709.1526756
21. Li JJ, Nenkova A (2015) Fast and accurate prediction of sentence specificity. In: Proceedings of 29th AAAI Conference on Artificial Intelligence, pp 2281–2287
22. Loper E, Bird S (2002) NLTK: The Natural Language Toolkit. In: Proceedings of ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, vol 1, pp 63–70. https://doi.org/10.3115/1118108.1118117
23. Maystre L (2017) Choix —Inference algorithms for models based on Luce's choice axiom. https://github.com/lucasmaystre/choix/
24. Merriam-Webster (2017) Merriam-Webster Online Dictionary. http://www.merriam-webster.com/

25. Microsoft (2016) Microsoft Azure Bing Search API. https://azure.microsoft.com/ja-jp/services/cognitive-services/search/
26. Miller GA (1995) WordNet: A lexical database for English, vol 38. https://doi.org/10.1145/219717.219748
27. Nagasawa Y, Nakamura K, Nitta N, Babaguchi N (2017) Effect of junk images on inter-concept distance measurement: Positive or negative? In: Advances in Multimedia Modeling: 23rd International Conference on Multimedia Modeling Procs., Springer, Lecture Notes in Computer Science, vol 10133, pp 173–184. https://doi.org/10.1007/978-3-319-51814-5_15
28. Nakamura K, Babaguchi N (2015) Inter-concept distance measurement with adaptively weighted multiple visual features. In: Computer Vision — ACCV 2014 Workshops. Springer, Lecture Notes in Computer Science, vol 9010, pp 56–70. https://doi.org/10.1007/978-3-319-16634-6_5
29. Oxford University Press (2017) OED Online. https://en.oxforddictionaries.com/
30. Paivio A, Yuille JC, Madigan SA (1968) Concreteness, imagery, and meaningfulness values for 925 nouns. J Exp Psychol 76(1):1–25
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
32. Samek W, Wiegand T, Mueller KR (2017) Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. Computer Research Repository arXiv:1708.08296
33. Smolik F, Kriz A (2015) The power of imageability: How the acquisition of inflected forms is facilitated in highly imageable verbs and nouns in Czech children. J First Lang 35(6):446–465. https://doi.org/10.1177/0142723715609228
34. Thurstone LL (1927) The method of paired comparisons for social values. J Abnorm Psychol 21(4):384–400
35. Yahoo (2005) Flickr. https://www.flickr.com/
36. Yanai K, Barnard K (2005) Image region entropy: A measure of "visualness" of Web images associated with one concept. In: Proceedings of 13th ACM Multimedia Conference, pp 419–422. https://doi.org/10.1145/1101149.1101241

**Marc A. Kastner** received his BSc and MSc in Computer Science from Braunschweig University of Technology in Braunschweig, Germany, in 2013 and 2016. He is currently studying towards his PhD in Informatics at the Graduate School of Informatics of Nagoya University, Japan. His research focuses on multimedia, language and vision and semantic gap problems.

**Ichiro Ide** received his BEng, MEng, and PhD from The University of Tokyo in 1994, 1996, and 2000, respectively. He became an Assistant Professor at the National Institute of Informatics, Japan in 2000. Since 2004, he has been an Associate Professor at Nagoya University. He was also a Visiting Associate Professor at National Institute of Informatics from 2004 to 2010, an Invited Professor at Institut de Recherche en Informartique et Systèmes Aléatoires (IRISA), France in 2005, 2006, and 2007, a Senior Visiting Researcher at ISLA, Instituut voor Informatica, Universiteit van Amsterdam from 2010 to 2011. His research interest ranges from the analysis and indexing to retargeting of multimedia contents, especially in large-scale broadcast video archives, mostly on news, cooking, and sports contents. He is a senior member of IEICE and IPS Japan, and a member of JSAI, IEEE, and ACM.

**Yasutomo Kawanishi** received his BEng and MEng degrees in Engineering and a PhD degree in Informatics from Kyoto University, Japan, in 2006,2008, and 2012, respectively. He became a Post Doctoral Fellow at Kyoto University, Japan in 2012. He moved to Nagoya University, Japan as an Designated Assistant Professor in 2014. Since 2015, he has been an Assistant Professor at Nagoya University, Japan. His research interests are Pedestrian-centric Vision, which includes Pedestrian Detection, Tracking, and Retrieval, for surveillance and in-vehicle videos. He received the best paper award from SPC2009, and Young Researcher Award from IEEE ITS Society Nagoya Chapter. He is a member of IEICE and IEEE.

**Takatsugu Hirayama** received the M.E. and D.E. degrees in Engineering Science from Osaka University in 2002 and 2005, respectively. From 2005 to 2011, he had been a Research Assistant Professor at the Graduate School of Informatics, Kyoto University. In 2011, he moved to the Graduate School of Information Science, Nagoya University. He had been an Assistant Professor from 2012 to 2014, a Designated Associate Professor from 2014 to 2017. He is currently a Designated Associate Professor at the Institutes of Innovation for Future Society, Nagoya University. His research interests include computer vision (face recognition, visual attention modeling, action recognition) and human-computer interaction (multi-modal interaction design, internal state estimation, interaction dynamics analysis). He is a member of IEICE, IPS Japan, ACM, and IEEE.

**Daisuke Deguchi** received his BEng and MEng degrees in Engineering and a PhD degree in Information Science from Nagoya University, Japan, in 2001,2003, and 2006, respectively. He became a Post Doctoral Fellow at Nagoya University, Japan in 2006. From 2008 to 2012, he had been an Assistant Professor at the Graduate School of Information Science, Nagoya University. From 2012, He had been an Associate Professor in Information Strategy Office, Nagoya University, Japan. He is working on the object detection, segmentation, recognition from videos, and their applications to ITS technologies, such as detection and recognition of traffic signs. He is a member of IEICE, IPS Japan and IEEE.

**Hiroshi Murase** received his BEng, MEng, and PhD degrees in Electrical Engineering from Nagoya University, Japan. In 1980 he joined the Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993 he was a visiting research scientist at Columbia University, New York. From 2003 he is a professor of Nagoya University, Japan. He was awarded the IEICE Shinohara Award in 1986, the Telecom System Award in 1992, the IEEE CVPR (Conference on Computer Vision and Pattern Recognition) Best Paper Award in 1994, the IPS Japan Yamashita Award in 1995, the IEEE ICRA (International Conference on Robotics and Automation) Best Video Award in 1996, the Takayanagi Memorial Award in 2001, the IEICE Achievement Award in 2002, and the Ministry Award from the Ministry of Education, Culture, Sports, Science and Technology in 2003. Dr. Murase is a Fellow of IEEE, IEICE, and IPS Japan.

## Affiliations

**Marc A. Kastner[1] · Ichiro Ide[1] · Yasutomo Kawanishi[1] · Takatsugu Hirayama[2] · Daisuke Deguchi[3] · Hiroshi Murase[1]**

Ichiro Ide
ide@i.nagoya-u.ac.jp

Yasutomo Kawanishi
kawanishi@i.nagoya-u.ac.jp

Takatsugu Hirayama
takatsugu.hirayama@nagoya-u.jp

Daisuke Deguchi
ddeguchi@nagoya-u.jp

Hiroshi Murase
murase@i.nagoya-u.ac.jp

1 Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601 Japan

2 Institute of Innovation for Future Society, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601 Japan

3 Information Strategy Office, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601 Japan