On the Quantification of the Mental Image of Visual Concepts for Multi-modal Applications

Marc A. Kastner[1,a)] Ichiro Ide[2,3] Yasutomo Kawanishi[3]
Takatsugu Hirayama[4] Daisuke Deguchi[3] Hiroshi Murase[3]

Abstract: The semantic gap is the lack of coincidence between the information one can extract from data and its interpretation. It is an yet solved issue for multimedia applications like image captioning, where it is often challenging to select the best fitting wording out of a group of candidates. To create a measurement for the perceived differences between words, this doctoral research proposes the idea of analyzing crawled image data to gain a better semantic understanding of language and vision. Abstract words have a broad mental image due to them being less visually defined, while concrete words with a rather narrow visual feature space are visually easier to grasp. The core goal of this research is to approximate this perceived abstractness of those words as a metric. The thesis proposes two methods, looking at both relative and absolute measurements. For the relative method, a data-driven approach is proposed, while the absolute measurements train a machine-learned model on existing data from Psycholinguistics. Each method is evaluated using crowd-sourced data and compared to related approaches. With this, the thesis presents methods to analyze the mental image of words from different angles, targeting a way to quantify the semantic gap between vision and language.

Keywords: Multimedia Modeling, Language and Vision, Computational Psycholinguistics, Visual Concept Semantics

1. Introduction

With the rise of multimedia applications including various modalities like text, images, and videos, the need for a better understanding of the semantic relationship between language and vision becomes more and more prominent. While the relationship between language and vision is part of various research, the perception for humans is often only of secondary consideration. For vision applications, the detection of concrete objects is a heavily researched field, but processing abstract ideas or concepts still easily causes issues, producing unnatural or unexpected results. A transfer of abstract ideas through modalities, e.g. from text to image, is an even harder problem, usually requiring extra information, and also knowledge on human expectation. A clearer modeling of how humans perceive language can improve word choice problems in various multimedia applications.

Through this doctoral research[30], the semantic gap between different concepts and words regarding human perception was studied. The central idea is to quantify the mental image of concepts, in order to find a measurable metric of the degree of the semantic gap between different concepts. The mental image in this context is considered as "the creation or re-creation of an experience generated from memorial information" [41]. Regarding a concept or word, the goal is to analyze whether its mental image is *clear*, e.g. it creates a rather concrete image and thus is easy to depict, or its mental image is *vague*, meaning that is is rather abstract, can mean various things, or is not clearly visually defined. For each input word, the methods' goal would be to estimate a value quantifying this idea on a scale. This overall goal is illustrated in Fig. 1.

The research is based on the core assumption that the average mental image regarding words across society is reflected in the images available through the Web and Social Media. As such, if we can gather a sufficient number of images related to a word or concept, the visual feature space will converge towards the average mental image of the said word or concept. Following, an exhaustive analysis of such images available through the Web and Social Media would yield meaningful information for estimating a metric predicting the mental image of a concept. This core assumption is illustrated in Fig. 2.

The doctoral thesis proposes two methods looking at this quantification from two angles: First, a relative measurement looking at granular differences of related concepts, and second, an absolute measurement looking at the gen-

[1] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
[2] Mathematical & Data Science Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
[3] Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
[4] Institute of Innovation for Future Society, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
[a)] mkastner@nii.ac.jp

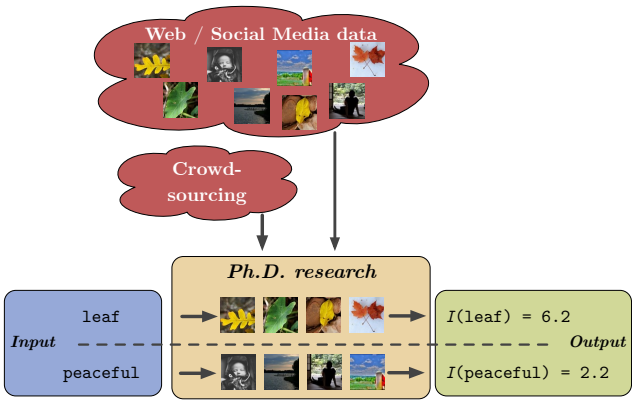


Fig. 1: Goal of this doctoral research. Input concepts or words are quantified regarding their human perception. As a source of information, images from Web and Social media are analyzed. Results are relative and/or absolute measurements describing the common mental image of the input word.

Fig. 2: Core assumption of this research. When asking a group of people about a word or concept, e.g., *peaceful*, the collective mental image across the mass of people would approximately converge to the images found when crawling the Web for the same word or concept.

eral trend across arbitrary concepts.

In the first work [28], the goal is the relative measurement of visual variety. The feature variety of images for concepts and their sub-concepts are compared to find granular differences regarding related concepts. In this way, it was possible to compare the visual gap between, e.g., *sports car* and *car* versus that between *car* and *vehicle*. The method chooses a data-driven approach, as the corpus of images for each concept has a strong influence on such comparisons. By reconfiguring super-ordinate concept datasets from their sub-ordinate concepts with a ratio that presumably reflects the occurrence of the subordinate concepts in real life, a dataset with less bias is created. As a ratio for the reconfiguration, various Web-based metrics were tested. For the experiments, a ground truth was obtained through a crowdsourced survey. The results showed that the visual variety estimation of less biased datasets resulted in a significant improvement of results.

In the second work [29], the goal is the absolute measurement of visual variety. Here, data-mining is applied on noisy Web-crawled images to estimate the *imageability* of a word. Imageability [49], a concept from Psycholinguistics, is a Likert-scale based metric quantifying words from low-imageability (e.g., *peace*) to high-imageability (e.g., *car*). In essence, the core idea was the observation that high-imageability words would commonly result datasets composed of very similar images, while usually less visually defined low-imageability words have a much higher variety. Using a variety of low- and high-level visual features and cross-similarity of related images, a model is trained to estimate an imageability score from a set of images related to a word. The method is compared to existing methods using textual data mining, finding that either modality excels in different areas.

In summary, this doctoral research targeted the semantic gap between language and vision from a less researched point of view. Incorporating ideas from Psycholinguistics, this research builds the foundation to bridge visual concept semantics with human perception. Looking at both relative and absolute measurements of the goal of mental image quantification, the research has the potential to be used for various applications.

Section 2 briefly overviews existing research related to the doctoral thesis. As the first proposed method, Sec. 3 discusses the relative visual variety measurements for concepts in a narrow domain. This is done using a data-driven approach comparing the clusters of visual features across different datasets. As the second proposed method, Sec. 4 discusses the absolute visual variety measurements for arbitrary concepts. In this approach, existing datasets are trained with ground-truth obtained from Psycholinguistics research to train a machine-learning model to estimate absolute visual variety scores. Before concluding in Sec. 6, Sec. 5 will discuss the differences of both approaches both in regarding their interpretation as well as in which applications they can be beneficial.

2. Related Work

There is various existing research related to the semantic gap and multi-modal applications which become relevant for this doctoral research. Furthermore, when looking at language and vision in terms of perception, there is some related research worth pointing out in the field of Psycholinguistics and Psychology.

2.1 Semantic gap

The *semantic gap* in content-based image retrieval received most attention through the work by Smeulders et al. [60], got furthermore refined in its meaning for applications and future challenges through Nack et al. [42] and Dorai and Venkatesh [17]. Discussing the definitions of the semantic gap and sensory gap, this survey paper discussed a variety of usage patterns of image retrieval.

Over time, there has been much research in narrowing or bridging the semantic gap for image retrieval and recommendation system purposes. As such, there has been related research on improving Web document retrieval [51][74],

content-based recommendation systems [6], and relatedness of recommendations in general [3][24][70].

2.2 Human perception and Psycholinguistics

Perception can be defined as "experience where the content reflects and is caused by an afferent physical stimulus". In contrast, *cognition* is defined as "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses" [48]. Both concepts are linked [4] and can also influence *how* we see things [62] as they share neural structures [15].

Paivio et al. [49] first proposed the concepts of *imageability*, *concreteness*, and *meaningfulness* as measurements for human perception of natural language. These concepts were further analyzed and refined [54], also in context of their implications for use of language [8][38][61]. Imageability also has a relationship to learning of language, and has been discussed for further use in research of dyslexia [26]. The relationship between imageability and text difficulty [57] and word understanding [19] were part of research.

There are a number of imageability dictionaries for English [11][53] and various other languages [59][72]. However, as dictionaries are commonly created by hand through user studies involving test subjects, the creation is labor-intensive.

2.3 Multimodal modeling

In Multimedia research, the analysis of visual concepts has been ground for multiple works. Some research has been looking at the visual distance of two visual concepts [43][44]. In other work, adjectives [71] and adjective-noun pairs [35] were compared based on entropy.

There are visual knowledge databases using unsupervised crawling based on visual differences [16]. Furthermore, there is also research in using deep networks to model cross-domain information between text and images [58][64][65].

In *visual diversification* [68], the idea is to increase the number of visually distinct result images in image retrieval methods.

There has been various approaches on automatically creating ontologies using visual features [23][31]. Other research uses ontologies to improve the results of single tags by going up or down the hierarchy to select better fitting words [45].

In natural language processing, there is some research using psycholinguistic measurements. Using text-based data-mining, there are methods to estimate imageability [37] and concreteness [5]. Computational linguistics can also profit from these metrics as a complement to sentiment embeddings [32], for improving image-text correspondences [21], and for improving machine translations [22].

There are various tasks in multimedia relevant to human perception of language and vision [7][10].

2.4 Applications

There is some multimedia research that uses psycholinguistic features as a supplementary source of information. Following, it was used to find comprehensible documents [63] and predict sentence specificity [36]. In efforts to use imageability to improve semantic relationships between image and text, it was used for standard image captioning [46][47] as well as slogans on posters [73] and Twitter [*1] images with text [69].

One interesting use-case could be Explainable AI (XAI) [56]. In XAI, the goal is to gain a better understanding of the operation of black-boxed AI models. As research on black-boxed models for object detection [20] shows, the actually trained information is often very different from what a human expects a model to learn. Following, a better language and vision understanding could benefit from a better understanding of trained models.

Another typical use-case for psycholinguistic features is sentiment- or emotion-analysis. There is various research on sentiment and emotion in multimedia applications [34], spanning visualization, datasets [27], and recognition techniques [25]. For sentiment evaluation, there are datasets such as LIWC [50] and Empath [18], which connect words and language to motivation, thoughts, emotions, and other sentiment-based numerical ratings. Sentiment and emotion research analyzes the human gap of multiple modalities regarding human perception. As such, it has become the topic of regular workshops affiliated with both Multimedia [55] and Natural Language Processing conferences [1].

3. Relative Visual Variety Differences for Concepts in a Narrow Domain

In order to bridge the semantic gap between related words, a logical first step would be to quantify their gap. For related concepts, one could look at sub-ordinate relationships between words like *sports car*, *car*, *motor vehicle*, and *vehicle*. All these concepts are logically related but also share a close visual resemblance, as the images of more concrete concepts (sports car) are sub-sets of the group of images for more abstract concepts (vehicle). A concept like *sports car* consists of just images of sports cars, while an image set of *vehicle* will contain images of sports cars, but also images of many other types of vehicles, like trucks or motorbikes.

The first proposed method [28] analyzes such a relative variety of visual features across images of concepts in a narrow domain. For each concept, a set of images is created, visually describing this concept. As more concrete concepts are visual subsets of more abstract ones, the method applies a bottom-up approach to create image sets for more abstract concepts. In order to adjust the ratio of images when creating more abstract image sets, a Web-based popularity metric is used. This ensures that there are fewer images of less relevant concepts. In essence, the thought is that images of, e.g., *tanks* would not contribute to the mental image of *vehicles* as much as images of *cars* would contribute. Following, the popularity metric would result in a higher number of

[*1] https://www.twitter.com/

images of *cars* than of *tanks* when creating a dataset for *vehicles*. Lastly, the visual feature space of each image set is analyzed to find the number of visual clusters. As a result, a more visually complex set of images would result in a higher number, while a visually rather defined set of images results in a low number. This output is used as a visual variety score in the evaluations, comparing the visual variety scores of various concepts between differently created datasets to a self-obtained ground-truth value.

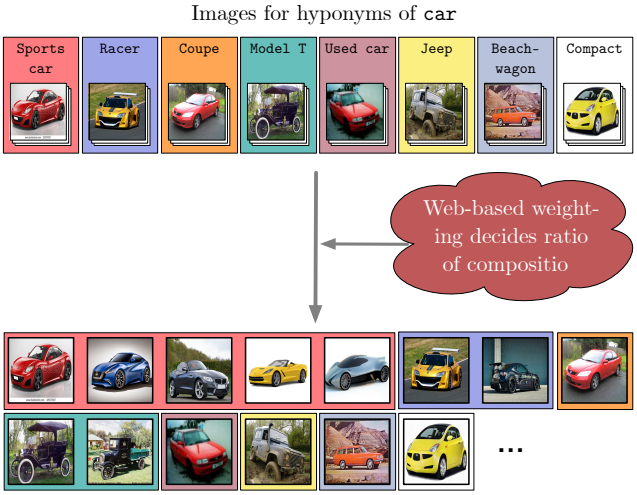
3.1 Dataset reconfiguration

Using WordNet [40] to obtain a hierarchy between words, the nature of existing image datasets for different words in the same tree were compared. Existing datasets like ImageNet [14] are often example-based. For each concept, as many example images describing this concept are collected as possible. In the example of ImageNet, this was done along the WordNet tree. Unfortunately, this results in rather unbalanced datasets, where some concepts have many related images and others none. Furthermore, it inherits existing biases coming from WordNet, which was not necessarily created with visual concepts in mind.

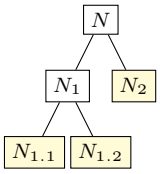
For this research, a well-composited dataset is needed to analyze and compare feature spaces across related datasets that become meaningful. We proposed a data-driven method, where the used image set for abstract concepts is a reconfiguration of its subordinate concept images. In a bottom-up fashion, the images for *car* are composited of images for *Sports car*, *Coupe*, and so on. To further influence the composition of images, a Web-based weighting decides the ratio of composition. This idea is illustrated in Fig. 3.

The dataset reconfiguration starts with a hierarchy of words extracted from WordNet. With this, each word has a hypernym-hypoynym relationship, from which one can infer their relationship. In a bottom-up fashion, the approach starts with all leaf nodes of the WordNet hierarchy. For those concepts, images are crawled from ImageNet and other Web sources like Search Engines [39]. For all non-leaf nodes, images of its hyponym concepts are combined to new image sets.

A core idea of the proposed method is the use of a Web-based weighting to decide the ratio of composition. For this, a popularity metric is determined using Web-based sources. The ratios of the popularity of concepts on the same level of the hierarchy are used to determine the number of images for each concept when merging them to a hypernym image set. For comparison, four different popularity metrics were tested: Using the Google Search API [*2], the number of search results for each concept is used as its popularity. The results for Google Image Search (GIS) and Google Text Search (GTS) were tested separately. Lastly, two word frequency-based methods were also analyzed, using corpora popular in computational linguistics such as Sketch Engine



(a) Idea of dataset reconfiguration.



(b) Getting hyponym leaf nodes of concept 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

(c) Determine weighting using a popularity metric.

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

(d) Recomposing the image corpus with p being the function for image retrieval.

Fig. 3: Dataset reconfiguration. Datasets for more abstract terms are created by compositing images from its hyponyms. A Web-based weighting decides the ratio of images during the reconfiguration process.

(SE) [33] and the Corpus of Contemporary American English (COCA) [13]. An example of the weightings for the concepts *truck* and *car* is shown in Table 1.

The full process of dataset reconfiguration is shown in Fig. 4. Apart from the baseline dataset which is the original ImageNet image composition, three new datasets were created for the experiments: First, in an equally weighted approach, all hyponym concepts are equally merged (e.g., a vehicle consists of the same amount of images of *car* and *tank*). Second and third, the Web popularity weighting obtained from Google Image Search and Google Text Search decides the ratio of hyponym concepts (e.g., the number of images of *car* and *tank* in vehicle is decided by the number of search results of each.) An example of resulting datasets is shown in Fig. 5.

In order for the visual feature spaces of datasets to be comparable, it is preferable to analyze the same number of

[*2] https://developers.google.com/custom-search/

Table 1: Comparison of different Web popularity measurements.

(a) Weighting for the concept *truck*

Leaf node	GTS	GIS	SE	COCA
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) Weighting for the concept *car*

Leaf node	GTS	GIS	SE	COCA
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%

images for each concept. Thus, during the crawling and composition steps, a random-sampling is done for concepts with a larger amount of images.

3.2 Visual variety measurements

In order to calculate a relative visual variety score, the visual feature space of each image set is analyzed and compared separately. Using all images available in the datasets (e.g., all images of all concepts), a Bag-of-Visual-Words (BoVW) model is trained using SURF descriptors [2][12]. Then, the visual feature space of a single concept is analyzed by extracting BoVW histograms of all its images. Using mean-shift clustering [9], the visual feature space is clustered for an open number of clusters decided by the algorithm. This is crucial, as in the core assumption a visually defined concept would yield in a very few number of clusters, while a rather vague, hard-to-grasp concept would yield in a very scattered feature space with many clusters. In the proposed method, the number of clusters in the visual feature space is calculated separately for each concept. This number of clusters is used as the resulting visual variety score in comparing related concepts.

For this approach to yield purposeful results, an existing *ideal* dataset with a meaningful composition of images is assumed, as discussed in the introduction of this section. For the evaluations, the dataset reconfiguration method explained above is used.

3.3 Experiments

The proposed method above was implemented and tested using four datasets for a selection of 25 words related to ve-

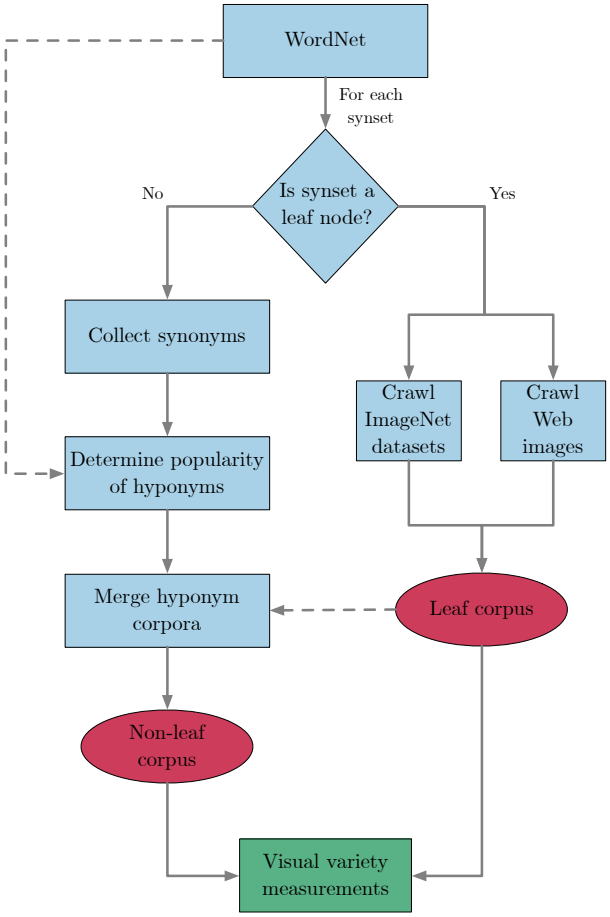


Fig. 4: Flowchart of the proposed data-driven method. For each leaf-node in a WordNet hierarchy, images are crawled. For each non-leaf node, a Web-based weighting of its hypoynym words is determined and then their subordinate concept images are merged to a new dataset for the hypernym word.

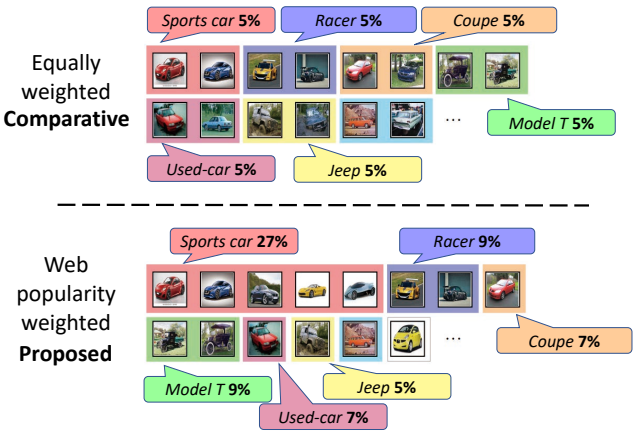


Fig. 5: Example of the reconfigured dataset. Each non-leaf node word is composited using its hyponym images. A ratio decides the composition. For a comparative method, the weighting for all merged concepts is set to be equal, while the proposed method uses a weighting determined by a Web-based popularity metric.

hicles. As a baseline method, ImageNet is used to analyze the bias of the existing dataset, using 1,000 images for each concept. For a comparative method, the ImageNet is also reconfigured using equal weighting, meaning that for the reconfiguration step, each hyponym is weighted equally, as shown in the upper part of Fig. 5. For the proposed method, datasets using the GTS and GIS as Web-based popularity metrics were tested. For the proposed methods, the reconfiguration allowed for crawling of additional images, resulting in 2,000 (GTS) and 2,430 (GIS) images being analyzed per concept. For each dataset, the visual variety score of each concept was calculated separately as outlined in the previous section.

In order to obtain a ground-truth annotation for a qualitative evaluation, a crowd-sourced survey was conducted on Social Media, including Facebook [*3], Twitter, and Reddit [*4]. In this study, 158 participants were asked about the variety of word pairs using Thurstone's paired comparison method [67]. After a comprehensive tutorial session, where each user is introduced to the idea of visual variety and taught how to approach each paired comparison in the main survey, the survey itself is conducted. The interface of the survey is shown in Fig. 6.

The results of all paired comparisons were used to obtain a human-verified ground-truth ranking of the visual variety. For the proposed method, the resulting number of clusters is normalized to the interval of [0, 100] between all its related terms, which yields a comparable ranking. For the evaluation, the ground-truth ranking is compared to the rankings estimated by the proposed method for each dataset.

The results are shown in Table 2. While a plain ImageNet did not correlate to the ground-truth labels obtained in the crowd-sourced survey, the dataset reconfiguration from the proposed method could improve results strongly. Each newly composited dataset can yield a better performance than the plain ImageNet dataset. This shows that ImageNet has an intrinsic bias when it comes to the composition of images across different subordinate concepts. It also showcases that a data-driven method can improve the results of a rather naive method for semantic analyses with only dataset changes. The equal weighting can already substantially improve both correlation and the mean squared error (MSE). When applying the GIS weighting as a Web-based popularity weighting, the results can be furthermore improved.

An example of the output is shown in Fig. 7, where the normalized relative visual variety scores for a selection of vehicles are illustrated.

3.4 Summary

In this research, a method to measure the relative visual variety of terms using reconfigured datasets modified to reflect Web-based popularity ratios was proposed. Web data is used to create and enhance an image set for each con-

*3 https://www.facebook.com/
*4 https://www.reddit.com/

(a) Survey: Tutorial

(b) Survey: Main part

Fig. 6: Ground-truth annotations used for the evaluations were obtained by conducting a crowd-sourced survey asking participants paired comparisons. Before taking part in the survey, a comprehensive tutorial (a) explains the idea of visual variety and what a participant is supposed to visualize in their mind when answering the survey. Then, the interface shown in (b) is used for paired comparisons.

Table 2: Qualitative analysis of the proposed method. The proposed method uses Google Text Search (GTS) and Google Image Search (GIS) as popularity metrics for weighting.

Corpus	Correlation (1 = best)	MSE (0 = best)
Plain ImageNet (Baseline)	0.25	10.54
Equal weighting (Comparative)	0.62	9.23
GTS weighting (Prop. 1)	0.56	14.89
GIS weighting (Prop. 2)	**0.73**	**9.01**

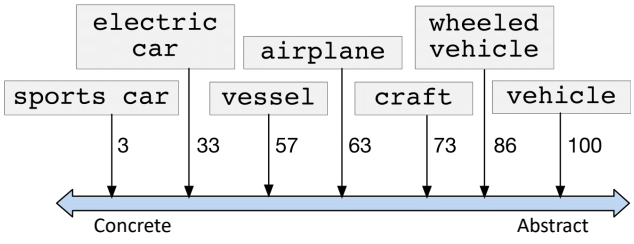


Fig. 7: Example of the relative visual variety scores for a selection of words related to vehicles. These results use the Google Image Search (GIS) weighting in the dataset reconfiguration step.

cept based on its popularity in social media. The method of counting the visual cluster calculates a distinct score 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 datasets with another, it shows that the correlation to ground truth highly depends on the used reconfiguration. Compared to the baseline corpus, the reconfiguration using the proposed method with a Google Image Search weighting improved the measurements by 192 percent, showing very promising results in terms of understanding the relationship between vision and language.

4. Absolute Visual Variety Estimation for Arbitrary Concepts

The human perception of words has been researched as part of Psycholinguistics since the 1960s [49]. In this research, metrics called *Imageability* or *Concreteness* express the ability for a human to (visually) imagine a word. These concepts are commonly described on a Likert scale from harder-to-imagine words like *peaceful* to easier-to-imagine words like *cat*. A limited scale typically does not allow for detailed comparisons like in relative measurements, but the relationship between imageability and the previously discussed visual variety is indisputable. On a global dictionary-level scale, its estimation comes with its problems. While it is easy to compare the number of clusters between, e.g., *vehicle* and *taxi*, an evaluation of *peace* and *cat* has no evident reference point for comparison.

The second proposed method discussed in this doctoral research [29] targets an absolute estimation of visual variety for arbitrary concepts. As such, imageability as a metric closely fits this goal. However, previous research commonly obtains ground-truth annotations for imageability using crowd-sourced surveys or user studies. A method that estimates such labels has been researched using text data-mining [37]. However, a visual approach using images has not been part of research, despite the evident connection of visual information in the form of images and the ability to visually imagine something as a mental image. To fix this gap, the proposed method analyzes the cross-similarity of Web-crawled images in order to train a model to estimate imageability as a metric for absolute visual variety.

In contrast to the work looking at relative visual variety estimation, this approach is mainly algorithm driven and does not assume a sophisticatedly created *ideal* dataset. This decision was initially out of convenience, as a data-driven method comes with its own bottlenecks through data acquisition processes and the amount of available data. Note that this second research method also targets an absolute measurements for an arbitrary domain. As such, there can not necessarily be a common WordNet hierarchy be assumed, as, e.g., *car* and *peace* do not share a meaningful hierarchy. As such, a bottom-up approach would not easily be applicable.

4.1 Imageability estimation

For a large number of images, a large set of related images is assumed. Using a selection of low- and high-level visual features is used to analyze the visual feature space of this concept. The visual features used for the experiments are discussed in Sec. 4.2. For each image and visual feature, a feature histogram is computed. Calculating the visual similarity between all images of the same concept, a similarity matrix is computed for each visual feature. The core assumption is similar to relative visual variety: A set of images for a high-imageability word would have somewhat similar images (e.g., all *cats* look similar), while a set of images for a low-imageability word is much more scattered (e.g., the concept *peace* is hard to visually define). From the obtained cross-similarity matrix, eigenvalues are computed. These eigenvalues intrinsically contain the trend of how visually similar a dataset is as individually encoded by each visual feature. The eigenvalues are used to train a regression model to predict the imageability of words. The proposed algorithm is outlined in Fig. 8.

4.2 Visual feature selection

In order to analyze each concept from multiple angles, a selection of features was chosen which look at visual characteristics from various viewpoints. Six visual features were chosen, which can logically be split into low-level and high-level features.

4.2.1 Low-level features

For lower-level features, the proposed method looks at traditional features used for CV pattern recognition.

First, the color distribution in the form of RGB color histograms was chosen to describe the overall mood of the images. In the context of imageability, this could also encode the atmosphere of an image, depending on e.g., colder or warmer colors. It also encodes information of visual adjectives like *yellow* or *bright*.

Second, global gradient descriptors in the form of GIST was chosen in order to describe global pattern distributions. Such features are commonly used for Web-retrieval engines.

Third, a local gradient descriptor in the form of Bag-of-Visual-Words (BoVW) [12] on SURF descriptorss [2] was chosen as a complementary feature. This type of feature is commonly used in traditional object detection, as it can decode which kind of patterns an object or scenes typically consist of.

4.2.2 High-level features

For high-level features, the proposed method uses pre-trained models, usually using deep learning, which approach image detection from a more *human* point of view. As such, rather than detecting raw edges, they are trained to encode human-annotated labels.

First, the theme of an image feature was chosen to encode the category of an image. This pre-trained model, rather than finding actual objects in an image, classifies an image into a category like *indoor*, *landscape*, or *architecture*. The setting of an image plays a large role in the similarity of

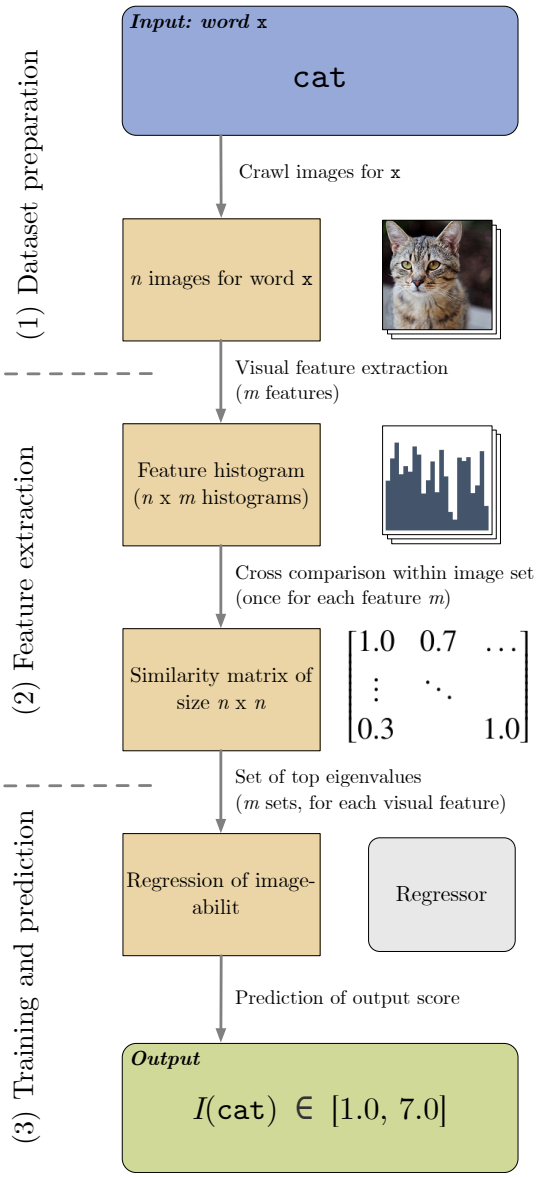


Fig. 8: Flowchart of the proposed method for absolute visual variety estimation. Using a large number of images for a concept, the visual feature space is analyzed using a variety of low- and high-level visual features. Cross-similarity of all images is encoded in a similarity matrix, whose eigenvalues give indication of the general trend of visual similarity across the dataset. The features obtained are used to train a regression model with an existing imageability dictionary as ground-truth.

images, as it is largely an encoding of backgrounds, which are often the largest part of each image in terms of surface area. Concretely, a pre-annotated annotation part of the YFCC100M dataset [66] consisting of 1,570 classes was employed.

Second, the contents of an image feature was chosen to further describe an image based on what it contains. Here, the object detection model YOLO [52] is used to detect all objects in each image. From this, a distribution histogram is computed which is used as a feature.

Third, the composition of image feature was chosen to encode where objects appear in images of a certain concept. Here, again, the object detection model YOLO is used, but rather than creating a distribution histogram, the spatial coordinates of detected objects are mapped to a 10×10 grid. The resulting feature histograms describe the location of objects found in each image, based on this grid.

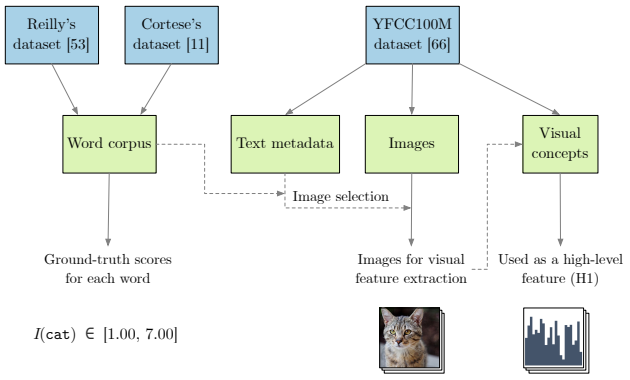


Fig. 9: Datasets used for this research. As imageability ground-truth annotations, two corpora from Psycholinguistics research are combined. The YFCC100M dataset is crawled for images and their text-annotations to obtain images for each concept.

4.3 Dataset construction

In this research, two types of datasets are employed. The first is a dictionary with (English language) words and their imageability annotations, which provides the ground truth for both the training process and the evaluation. The second is a large number of images for each word, which will be used for visual feature extraction. The relationship between both types of datasets is illustrated in Fig. 9.

4.3.1 Imageability dictionary

There are a number of imageability dictionaries across various languages, including English [11][53], Indonesian [59], and Cantonese [72]. As described before, imageability dictionaries try to quantify the human perception of words, usually averaging the results over all test subjects. Low-imageability words would be things where one can not, or only hardly, grasp a mental image to describe it. These are usually rather abstract concepts, like *peace* or the word *abstract* itself. High-imageability words, on the other hand, are something rather concrete, usually easier to visually grasp, like *dog* or the color *red*.

Due to their nature, such dictionaries are commonly created by hand. Using crowd-sourcing or surveys, a set of words is judged by each test subject. Due to the amount of labor involved in this process, most dictionaries are rather small.

For this research, English imageability dictionaries by Reilly et al. [53] and Cortese et al. [11] in combination are used as a ground-truth for imageability. These datasets provide the results as a Likert scale score averaged over all test subjects, in the range of [1.0, 7.0]. While the former is only composed of nouns, the latter includes other parts-of-speech.

In the case of overlap, the average of both dictionaries is taken.

While Likert scales are very common in Psychology, Computer Science is used to either percentual results or a normalized scale of [0.0, 1.0]. Therefore, for pure understandability of the evaluation results, the interval of [1.00, 7.00] is normalized to [0, 100] here.

4.3.2 Imagesets

Using the imageability dictionary as a basis, a large number of images for each word with imageability annotation are crawled from Social Media. For this report, the YFCC100M [66] dataset is employed. YFCC100M is based on the US photography social media platform Flickr [*5]. The noisy Web-based origin ensures a composition that comes close to how a human perceives the concept. The dataset consists of 100 million images posted to Flickr up to 2014, each annotated with some text-based annotations like a title, a description, user taggings, and more. The text-based annotations are used to identify the relationship between images and words.

If a word from the imageability dictionary is contained in one of the text-based annotations (title, description, or user-tagging) of an image, it is considered to be related to this word. For example, if the title of an image is "This is my peaceful cat", the image would become part of the data for the concepts *peaceful* and *cat*. An equal number of images is crawled for every word, which makes sure that the similarity matrices of all concepts are comparable. A high number of images is preferable to average out noise and selection bias. As a trade-off between crawling time, processing time, and accuracy, 5,000 images were obtained for each word.

4.4 Experiments

For the evaluation of the proposed method, a dataset consisting of 5,000 images for 582 words with imageability annotations was constructred. The proposed method uses a combination of all six visual features described above. For comparison, the single features were tested separately. Furthermore, the text-based data-mining method by Ljubesic [37] was used as a comparative method, along with our relative visual variety method described in the previous section.

Table 3 shows the qualitative results. As shown, the results for text data-mining and image data-mining can yield comparable results. While the text data-mining has a better correlation, the image data-mining gets slightly better results for the mean absolute error (MAE). Looking at the single visual features, we can see that they can each complement one another, resulting in the best results when combining all of them.

When digging further, one can look at which set of features works better for low-imageability words vs. high-imageability ones. High-level features were trained on neural networks for object detection and could thus excel in estimating the imageability of more concrete words, as these images commonly contain actual detectable objects. In contrast, low-level features capture the overall atmosphere of a concept better, thus they have the potential of yielding better results for abstract concepts like *peace* or *art*, where there are not necessarily shared objects across multiple images. Table 4 shows an analysis of this idea. As assumed, high-level features excel for high-imageability words, while low-level features yield better results for low-imageability words.

Figure 10 shows the testing dataset visualized as a scatter plot. We can see that the proposed method has an advantage for low-imageability words, while the text-based method fits the ground-truth better for high-imageability ones. Table 5 shows some predicted examples and their corresponding ground-truths. They are calculated back to the interval of [1.0, 7.0] in order to better compare them to the ground-truth values. Lastly, Fig. 11 shows some results with example images in order to present a feeling on how the dataset relates to the predicted values.

Table 3: Qualitative analysis for different sets of visual features. Each set uses a different selection of visual features for the estimation. Comparative method 1 tries the proposed method of relative visual variety discussed in Sec. 3 for absolute estimation. Comparative method 2 uses text data mining instead of image data mining.

Feature	Correlation (1 : best)	MAE (0 : best)
L1: Color histogram	0.53	11.30
L2: SURF + BoVW	0.54	11.48
L3: GIST	0.42	12.05
H1: Image theme	0.62	10.19
H2: Image content	0.43	12.55
H3: Image composition	0.25	13.98
L* (Low only)	0.60	11.03
H* (High only)	0.61	10.18
All (Prop.)	0.63	**10.14**
Comp. 1 (Relative visual variety)	−0.01	67.31
Comp. 2 (Text data mining [37])	**0.70**	10.39

Table 4: Feature comparison for abstract words vs. concrete words. For this evaluation, the testing dataset has been split half at the median value. For MAE, the lower is the better, while for the correlation, 1 is the best.

Features		Low-imag.		High-imag.	
		Corr.	MAE	Corr.	MAE
Low	L1 (Color)	0.32	11.36	0.00	11.25
	L2 (SURF)	**0.36**	11.26	0.18	11.71
	L3 (GIST)	0.20	12.18	**0.20**	12.82
High	H1 (Theme)	0.26	11.37	0.19	9.32
	H2 (Content)	0.11	12.41	0.10	12.69
	H3 (Comp.)	−0.01	13.99	−0.05	13.87
Multiple	L* (Low)	0.32	**10.90**	0.16	11.37
	H* (High)	0.27	11.31	0.10	9.10
	All (Prop.)	0.26	10.79	0.17	10.11
Other	Text [37]	**0.40**	13.27	0.18	**7.51**

4.5 Summary

In this research, a method using image-based data mining with a variety of low-level and high-level visual features to estimate imageability scores for words was proposed. In

[*5] https://www.flickr.com/

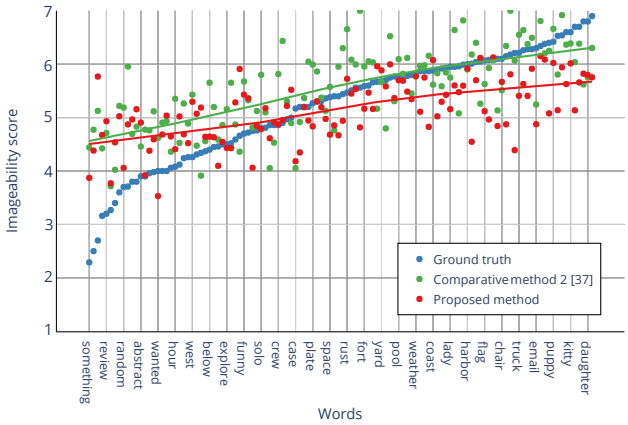


Fig. 10: Scatter plot of the results. The proposed method can more accurately estimate values for low-imageability concepts, while the comparative method using text data mining has a higher accuracy for high-imageability concepts.

Table 5: Imageability prediction results of the proposed method. The printed values are calculated back to the interval of [1.0, 7.0] in order to make them comparable to the Lickert scale of existing imageability dictionaries.

Imageability	Word	Predicted	Ground-truth
High	breakfast	5.91	6.28
	leaf	6.13	6.07
	plant	6.12	6.05
	coast	6.07	5.88
	pool	5.70	5.77
Low	early	3.90	3.91
	random	4.05	3.70
	challenge	4.38	3.96
	need	3.77	3.26
	break	4.59	3.97
Outliers	fauna	5.77	2.70
	review	3.19	4.93
	silver	4.39	6.20
	email	4.87	6.30
	plastic	5.07	6.40

previous related research, similar dictionaries have been created by hand through user studies or crowd-sourcing. This labor-intensive process results in a limited number of data samples compared to the full word corpora of languages. Using the proposed method, such an absolute metric can be estimated using data mining. The evaluations show an MAE of 10.14 (approximately 0.71 scores on the Lickert scale) and a correlation of 0.63 for the best feature combination. This shows that the results correlate to the ground-truth Lickert scale, especially as the error is less than one level on the Lickert scale. This performance could be considered enough for many applications, as the general trend of high-imageability versus low-imageability is sufficently captured. Furthermore, the evaluations give us an insight on which features excel for which type of words. In a general trend, the low-level features worked better for abstract words, while the high-level features worked better for concrete words. This is due to concrete terms often being related to objects, while abstract terms can only be estimated by encoding the general visual trends of atmosphere, gradients, and dataset noise. The proposed method is intended to be used to expand the vocabulary in imageability dictionaries. There are also opportunities to integrate them into multimodal applications like sentiment analyses. Another possible application that comes to mind is the quality assessment of auto-generated image captioning results. There, results could be assessed differently, depending on whether they are used for complementary information, accessibility purposes, or other use-cases.

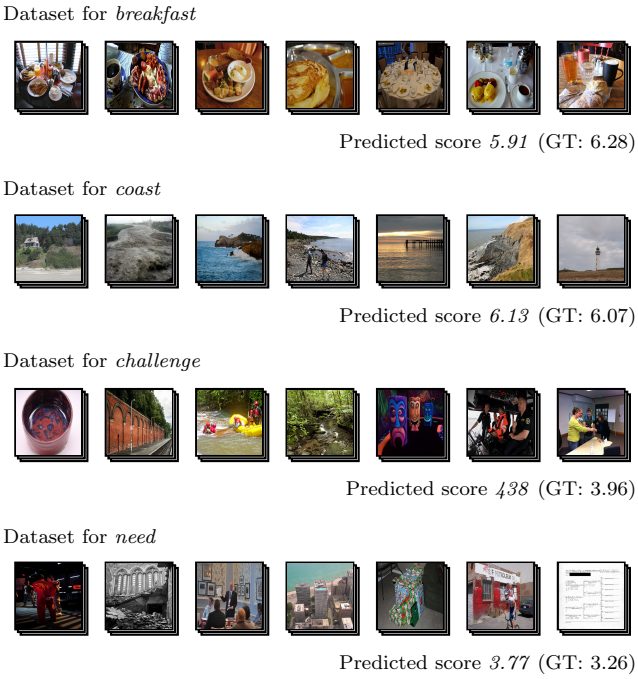


Fig. 11: Example of image datasets and their predicted imageability scores. Scores are in an interval of [1.0, 7.0] to compare them to the Likert scale of the ground-truth dataset.

5. Discussion

In this section, the doctoral research as a whole is discussed. First, there are some words on the core assumption discussed in the motivational introduction, and how the experimental results of the proposed methods can prove it, and which remaining problems exist with the core assumption. Second, the different applications for relative measurements vs. absolute measurements are discussed. Third, the advantages and disadvantages of the data-driven and algorithm-driven methods are outlined. Lastly, there are some comments on the reproducibility of the proposed research.

5.1 Core assumptions

During the motivation of this research, it was assumed that the average mental image regarding words across society would be reflected in the images available through the Web and Social Media. Following this idea, the proposed research applied data mining on visual characteristics of Web images to estimate visual variety and imageability. The results show that this assumption holds true for the

chosen datasets; for both absolute and relative measurements, the experiments showed promising results, closely resembling the expectations available through ground-truth annotations made by humans.

One thing that the contributed methods could not consider is personalized scores. This is due to the core assumption discussed before focusing on the average mental image across society. For the relative measurements, a personalized score could be targeted by modifying the popularity metric towards personalized popularity. For the absolute measurements, the idea of personalization, unfortunately, goes somewhat against the general concept of imageability being an average of multiple asked people.

Another thing worth mentioning is the role of dataset bias. Noise is an issue this research is only secondarily concerned with. Mainly, the core assumption intrinsically assumes noise for more abstract concepts. Crawling for *peace* will naturally result in much more noisy images than crawling for *cat*. However, both methods, looking at similarity within the dataset, already adjust for that, as a high noise ratio will commonly be assumed to be something vague and thus having a high variety. A Web-bias, on the other hand, is something both methods might struggle with. Terms that are often used in marketing might be over-represented in Web searches compared to their expectations in daily life. For example, the word *sports car* resulted in overly high popularity across all metrics, which might be due to Web-related bias. Another interesting outlier was the word *canon*, which is overly represented in queries on photography websites. This is due to the proper noun *Canon* (as in the camera brand) rather than any other meaning of the word.

5.2 Relative measurement vs. absolute measurement

In this report, two methods to approach mental image quantification for absolute and relative comparison were discussed. While both methods measure different targets, neither can be considered *better*. Following, the usefulness of the results would depend on the application.

The method for relative visual variety has been tested for a group of concepts related to vehicles. Similarly, it is expected to work as well for a narrow domain like *animals*, *plants*, and so on. For applications like image tagging, such granular semantic differences could prove beneficial for word-selection tasks. While an image of a car could be tagged with *vehicle*, *car*, *sports car*, or *Car-brand Model-XXXX*, depending on the application, some might be too verbose while others might be too vague. In such a task, a direct comparison of the semantics of candidates might be useful, where the method for relative visual variety can find minute differences between close candidates.

In contrast, the approach for absolute measurements looks at the big picture. While *car* and *sports car* might even yield almost the same level on the Likert scale, it would be useful in estimating the general trend of a word. For example, when estimating text difficulty, understandability or visualness of a caption, it would be interesting to find the overall trend (e.g., are the words used too vague or too concrete).

Lastly, one could combine both approaches: First using the absolute measurement to find the general trend, and then use the relative measurement to decide between similar candidates for the same level.

5.3 Data-driven vs. algorithm-driven approaches

For each proposed method, different approaches were applied to solve the sub-problem: The former uses a data-driven method, while the latter uses an algorithm-driven approach.

For relative measurements, the assumption was that the ratio of such sub-concepts relates to how humans create a mental image of the parent concept, as a sub-concept daily seen in daily life (e.g., *car*) may have a stronger influence than a concept rarely seen (e.g., *jet*). Following, a reconfiguration of the datasets to decrease bias seemed natural and also led to promising results. However, it comes with some downsides: First, the number of images available for some concepts often heavily bottleneck other ones when it comes to the reconfiguration. Second, as the approach is tied to WordNet [40] and ImageNet [14], it only works for nouns which are available in both datasets. Third, using a proprietary API for both the image crawling and obtaining the Web-based popularity measurement results in unnecessary cost, but also issues due to the black-boxed nature of those proprietary APIs.

In contrast, for absolute measurements, an algorithm-driven approach was chosen. Due to training and testing data being available through imageability dictionaries, one could train a model on existing image data. One problem found through this is the clutteredness of the original data. With the ground-truth Likert scale being an average of all test subjects asked, most concrete words between levels 5 and 7 are simply clustered together with no obvious or meaningful up-down relationship. This makes training a good correlation very difficult as most granular differences are even incomprehensible for the human. Here, the hierarchical data-driven approach had its advantages, as the composite nature would force a clear ranking of granularly different concepts.

5.4 Reproducibility of published work

In recent years, the reproducibility of academic results has become more and more focused on in the research community.

For the relative measurements, while the actual implementation of the data-driven method would be trivial, the interesting thing is the used and created dataset. However, due to copyright reasons, crawled images can not be republished easily. One source of concern is the black-boxed nature of the APIs for dataset retrieval. It is also questionable whether redoing the experiments would yield the exact same results, as every crawling might yield a slightly different set of images due to updated indices in the APIs used for dataset

retrieval. The same is true for the popularity metrics. A potential advantage of the method, on the other hand, could be that it naturally adjusts to changes over time, as recrawling the data also *updates* it regarding to how popularity might have changed over time.

For the absolute measurements, the source code, as well as pre-trained models, have been made available on GitHub [*6]. Additionally, students in the lab have been using this framework for their own research, yielding similar results using the same as well as separately crawled datasets.

6. Conclusion

The research described in the doctoral thesis attempts to quantify the perceived variety of concepts from a visual standpoint, in order to get a better understanding of semantic distances. An insufficient understanding of language and vision can result in word choice problems in image captioning or machine translations, among other problems. To tackle this research, the main problem of mental image quantification was divided into two sub-tasks, which were proposed and tested; Relative and absolute measurements. For the first topic, in order to measure the relative visual variety of concepts in a narrow domain, datasets for related concepts were reconfigured based on their hyponym concepts in a data-driven approach. For the second topic, in order to measure the absolute visual variety of arbitrary concepts, images for general-purpose words were crawled from Social Media and analyzed towards their visual characteristics, in order to train a model to estimate their imageability.

Both methods were evaluated and compared to other related methods, where either method showed promising results. Following, looking at and tackling the afore-motivated problem from two different angles, this solves the aim of this doctoral research. Combined, these two research topics provide a contribution to the challenging field of understanding the semantic gap between vision and language. Both evaluations established the comparison of Web-crawled image datasets as a viable method for analyzing the perceived variety of concepts. This gives further insights for word choice problems and other tasks. The methods have the potential to serve as a metric and source of knowledge for the perceived differences between concepts, connecting visual data and language for multimodal modeling.

As a remaining challenge, a combination of both methods would be interesting. While both relative and absolute measurements have their purposes, for many applications, a two-step approach applying both ideas would be thinkable. This could give even better results tailoring the idea of visual variety closer to many applications. Furthermore, especially the evaluations of the second method for absolute measurements showed that text-based data-mining and image-based data-mining excel in different fields. As a result, a multi-modal approach mining both modalities would be a logical next step for research. Lastly, in personal interest, a cross-language analysis of these ideas would be meaningful in order to compare results across different cultures or interest groups.

Acknowledgments Parts of this research were supported by JSPS KAKENHI 16H02846, and Microsoft Research CORE16 joint research project.

[*6] https://github.com/mkasu/imageabilityestimation/

References

[1] Balahur, A., Mohammad, S. M., Hoste, V. and Klinger, R.(eds.): *Proc. 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium, Association for Computational Linguistics (2018).

[2] Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L.: Speeded-Up Robust Features (SURF), *Comput. Vis. Image Underst.*, Vol. 110, No. 3, pp. 346–359 (online), DOI: 10.1016/j.cviu.2007.09.014 (2008).

[3] Budanitsky, A. and Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, *Proc. 2nd Meeting of the North American Chapter of the Association for Computatonal Linguistics*, pp. 29–34 (2001).

[4] Cahen, A. and Tacca, M. C.: Linking perception and cognition, *Front. Psychol.*, Vol. 4, p. 144 (online), DOI: 10.3389/fpsyg.2013.00144 (2013).

[5] Charbonnier, J. and Wartena, C.: Predicting Word Concreteness and Imagery, *Proc. 13th Int. Conf. on Computational Semantics*, pp. 176–187 (2019).

[6] Cheng, S., Chen, W. and Sundaram, H.: Semantic visual templates: Linking visual features to semantics, *Proc. 1998 Int. Conf. on Image Processing*, Vol. 3, pp. 531–535 (online), DOI: 10.1109/ICIP.1998.727321 (1998).

[7] Cohendet, R., Demarty, C., Duong, N. Q. K., Sjöberg, M., Ionescu, B., Do, T. and Rennes, F.: MediaEval 2018: Predicting Media Memorability Task, *Computing Research Repository*, No. arXiv:1807.01052 (2018).

[8] Coltheart, V., Laxon, V. J. and Keating, C.: Effects of word imageability and age of acquisition on children's reading, *Br. J. Psychol.*, Vol. 79, No. 1, pp. 1–12 (online), DOI: 10.1111/j.2044-8295.1988.tb02270.x (1988).

[9] Comaniciu, D. and Meer, P.: Mean Shift: A robust approach toward feature space analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 24, No. 5, pp. 603–619 (online), DOI: 10.1109/34.1000236 (2002).

[10] Constantin, M. G., Redi, M., Zen, G. and Ionescu, B.: Computational understanding of visual interestingness beyond semantics: Literature survey and analysis of covariates, *ACM Comput. Surv.*, Vol. 52, No. 2, pp. 25:1–25:37 (online), DOI: 10.1145/3301299 (2019).

[11] Cortese, M. J. and Fugett, A.: Imageability ratings for 3,000 monosyllabic words, *Behav. Res. Methods Instrum. Comput.*, Vol. 36, No. 3, pp. 384–387 (online), DOI: 10.3758/BF03195585 (2004).

[12] Csurka, G., Dance, C. R., Fan, L., Willamowski, J. and Bray, C.: Visual categorization with bags of keypoints, *Proc. ECCV 2004 Workshop on Statistical Learning in Computer Vision*, pp. 1–22 (2004).

[13] Davies, M.: The corpus of contemporary American English: 520 million words, 1990–present. Harvard Dataverse. Available online at http://corpus.byu.edu/coca/. (2008).

[14] Deng, J. D. J., Dong, W. D. W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, *Proc. 2009 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 2–9 (online), DOI: 10.1109/CVPR.2009.5206848 (2009).

[15] Dijkstra, N., Bosch, S. E. and van Gerven, M. A. J.: Shared neural mechanisms of visual perception and imagery, *Trends Cogn. Sci. (Regul. Ed.)*, Vol. 23, No. 5, pp. 423–434 (2019).

[16] Divvala, S. K., Farhadi, A. and Guestrin, C.: Learning everything about anything: Webly-supervised visual concept learning, *Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3270–3277 (online), DOI: 10.1109/CVPR.2014.412 (2014).

[17] Dorai, C. and Venkatesh, S.: Bridging the semantic gap with computational media aesthetics, *IEEE MultiMed.*, Vol. 10, No. 2, pp. 15–17 (online), DOI: 10.1109/MMUL.2003.1195157 (2003).

[18] Fast, E., Chen, B. and Bernstein, M. S.: Empath: Understanding topic signals in large-scale text, *Computing Research Repository*, No. arXiv:1602.06979 (2016).
[19] Giesbrecht, B., Camblin, C. C. and Swaab, T. Y.: Separable effects of semantic priming and imageability on word processing in human cortex, *Cereb Cortex*, Vol. 14, No. 5, pp. 521–529 (online), DOI: 10.1093/cercor/bhh014 (2004).
[20] Hentschel, C. and Sack, H.: What image classifiers really See —Visualizing bag-of-visual words models, *Advances in Multimedia Modeling: 21st Int. Conf. on Multimedia Modeling Procs.*, Lecture Notes in Computer Science, Vol. 8935, Springer, pp. 95–104 (online), DOI: 10.1007/978-3-319-14445-0_9 (2015).
[21] Hessel, J., Mimno, D. and Lee, L.: Quantifying the visual concreteness of words and topics in multimodal datasets, *Proc. 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1*, pp. 2194–2205 (online), DOI: 10.18653/v1/N18-1199 (2018).
[22] Hewitt, J., Ippolito, D., Callahan, B., Kriz, R., Wijaya, D. T. and Callison-Burch, C.: Learning translations via images with a massively multilingual image dataset, *Proc. 56th Annual Meeting of the Association for Computational Linguistics, vol. 1*, pp. 2566–2576 (online), DOI: 10.18653/v1/P18-1239 (2018).
[23] Inoue, N. and Shinoda, K.: Adaptation of word vectors using tree structure for visual semantics, *Proc. 24th ACM Multimedia Conf.*, pp. 277–281 (online), DOI: 10.1145/2964284.2967226 (2016).
[24] Jiang, J. J. and Conrath, D. W.: Semantic similarity based on corpus statistics and lexical taxonomy, *Proc. 10th Research on Computational Linguistics Int. Conf.*, pp. 19–33 (1997).
[25] Jindal, S. and Singh, S.: Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning, *Proc. 2015 IEEE Int. Conf. on Image Processing*, pp. 447–451 (online), DOI: 10.1109/INFOP.2015.7489424 (2015).
[26] Jones, G. V.: Deep dyslexia, imageability, and ease of predication, *Brain. Lang.*, Vol. 24, No. 1, pp. 1–19 (online), DOI: 10.1016/0093-934X(85)90094-X (1985).
[27] Jou, B., Chen, T., Pappas, N., Redi, M., Topkara, M. and Chang, S.: Visual affect around the world: A large-scale multilingual visual sentiment ontology, *Computing Research Repository*, No. arXiv:1508.03868 (2015).
[28] Kastner, M. A., Ide, I., Kawanishi, Y., Hirayama, T., Deguchi, D. and Murase, H.: Estimating the visual variety of concepts by referring to Web popularity, *Multimed. Tools Appl.*, Vol. 78, No. 7, pp. 9463–9488 (online), DOI: 10.1007/s11042-018-6528-x (2019).
[29] Kastner, M. A., Ide, I., Nack, F., Kawanishi, Y., Hirayama, T., Deguchi, D. and Murase, H.: Estimating the imageability of words by mining visual characteristics from crawled image data, *Multimed. Tools Appl.*, (online), DOI: 10.1007/s11042-019-08571-4 (2020).
[30] Kastner, M. A.: Quantifying the mental image of visual concepts, PhD Thesis, Nagoya University (2020).
[31] Kawakubo, H., Akima, Y. and Yanai, K.: Automatic construction of a folksonomy-based visual ontology, *Proc. 2010 IEEE Int. Symposium on Multimedia*, pp. 330–335 (online), DOI: 10.1109/ISM.2010.57 (2010).
[32] Kiela, D., Hill, F., Korhonen, A. and Clark, S.: Improving multi-modal representations using image dispersion: Why less is sometimes more, *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 835–841 (online), DOI: 10.3115/v1/P14-2135 (2014).
[33] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovávr, V., Michelfeit, J., Rychlý, P. and Suchomel, V.: The sketch engine: Ten years on, *Lexicography*, Vol. 1, No. 1, pp. 7–36 (online), DOI: 10.1007/s40607-014-0009-9 (2014).
[34] Kim, E. and Klinger, R.: A survey on sentiment and emotion analysis for computational literary studies, *Computing Research Repository*, No. arXiv:1808.03137 (2018).
[35] Kohara, Y. and Yanai, K.: Visual analysis of tag cooccurrence on nouns and adjectives, *Advances in Multimedia Modeling: 19th Int. Conf. on Multimedia Modeling Procs.*, Lecture Notes in Computer Science, Vol. 7732, Springer, pp. 47–57 (online), DOI: 10.1007/978-3-642-35725-1-5 (2013).
[36] Li, J. J. and Nenkova, A.: Fast and accurate prediction of sentence specificity, *Proc. 29th AAAI Conf. on Artificial Intelligence*, pp. 2281–2287 (2015).
[37] Ljubešić, N., Fišer, D. and Peti-Stantić, A.: Predicting concreteness and imageability of words within and across languages via word embeddings, *Proc. 3rd Workshop on Representation Learning for NLP*, pp. 217–222 (online), DOI: 10.18653/v1/W18-3028 (2018).
[38] Ma, W., Golinkoff, R. M., Hirsh-Pasek, K., McDonough, C. and Tardif, T.: Imageability predicts the age of acquisition of verbs in Chinese children, *J. Child. Lang.*, Vol. 36, pp. 405–423 (online), DOI: 10.1017/S0305000908009008 (2009).
[39] Microsoft: Microsoft Azure Bing Search API (2016).
[40] Miller, G. A.: WordNet: A lexical database for English, *Comm. ACM*, Vol. 38, No. 11, pp. 39–41 (online), DOI: 10.1145/219717.219748 (1995).
[41] Morris, T., Spittle, M. and Watt, A. P.: *Imagery in sport*, Human Kinetics (2005).
[42] Nack, F., Dorai, C. and Venkatesh, S.: Computational media aesthetics: Finding meaning beautiful, *IEEE MultiMed.*, Vol. 8, No. 4, pp. 10–12 (online), DOI: 10.1109/93.959093 (2001).
[43] Nagasawa, Y., Nakamura, K., Nitta, N. and Babaguchi, N.: Effect of junk images on inter-concept distance measurement: Positive or negative?, *Advances in Multimedia Modeling: 23rd Int. Conf. on Multimedia Modeling Procs.*, Lecture Notes in Computer Science, Vol. 10133, Springer, pp. 173–184 (online), DOI: 10.1007/978-3-319-51814-5_15 (2017).
[44] Nakamura, K. and Babaguchi, N.: Inter-concept distance measurement with adaptively weighted multiple visual features, *Computer Vision —ACCV 2014 Workshops*, Lecture Notes in Computer Science, Vol. 9010, Springer, pp. 56–70 (online), DOI: 10.1007/978-3-319-16634-6_5 (2015).
[45] Ordonez, V., Deng, J., Choi, Y., Berg, A. C. and Berg, T. L.: From large scale image categorization to entry-level categories, *Proc. 2013 IEEE Int. Conf. on Computer Vision*, pp. 2768–2775 (online), DOI: 10.1109/ICCV.2013.344 (2013).
[46] Otto, C., Holzki, S. and Ewerth, R.: "Is this an example image?" —Predicting the relative abstractness level of image and text, *Computing Research Repository*, Vol. arXiv:1901.07878 (2019).
[47] Otto, C., Springstein, M., Anand, A. and Ewerth, R.: Understanding, categorizing and predicting semantic image-text relations, *Proc. 2019 Int. Conf. on Multimedia Retrieval*, pp. 168–176 (online), DOI: 10.1145/3323873.3325049 (2019).
[48] Oxford University Press: OED Online (2017).
[49] Paivio, A., Yuille, J. C. and Madigan, S. A.: Concreteness, imagery, and meaningfulness values for 925 nouns, *J. Exp. Psychol.*, Vol. 76, No. 1, pp. 1–25 (1968).
[50] Pennebaker, J. W., Francis, M. E. and Booth, R. J.: *Linguistic Inquiry and Word Count (LIWC): LIWC2001*, Erlbaum, Mahwah, NJ, USA (2001).
[51] R. Zhao and Grosky, W. I.: Narrowing the semantic gap —Improved text-based Web document retrieval using visual features, *IEEE Trans. Multimed.*, Vol. 4, No. 2, pp. 189–200 (online), DOI: 10.1109/TMM.2002.1017733 (2002).
[52] Redmon, J. and Farhadi, A.: YOLO9000: Better, faster, stronger, *Computing Research Repository*, No. arXiv:1612.08242 (2016).
[53] Reilly, J. and Kean, J.: Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications, *J. Cogn. Sci.*, Vol. 31, No. 1, pp. 157–168 (online), DOI: 10.1080/03640210709336988 (2010).
[54] Richardson, J. T. E.: Imageability and concreteness, *Bull. Psychol. Soc.*, Vol. 7, No. 5, pp. 429–431 (online), DOI: 10.3758/BF03337237 (1976).
[55] Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., Ciftçi, E., Güleç, H., Salah, A. A. and Pantic, M.(eds.): *Proc. 2018 Audio/Visual Emotion Challenge and Workshop*, Assoc. Comp. Mach., New York, NY, USA (2018).
[56] Samek, W., Wiegand, T. and Müller, K.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, *Computing Research Repository*, No. arXiv:1708.08296 (2017).
[57] Schwanenflugel, P. J.: Why are abstract concepts hard to understand?, *The Psychology of Word Meanings*, Psychology Press, New York, NY, USA, pp. 235–262 (2013).
[58] Shu, X., Qi, G.-J., Tang, J. and Wang, J.: Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation, *Proc. 23rd ACM Multimedia Conf.*, pp. 35–44 (online), DOI: 10.1145/2733373.2806216 (2015).
[59] Sianipar, A., van Groenestijn, P. and Dijkstra, T.: Affective

meaning, concreteness, and subjective frequency norms for Indonesian words, *Front. Psychol.*, Vol. 7, p. 1907 (online), DOI: 10.3389/fpsyg.2016.01907 (2016).

[60] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. and Jain, R.: Content-based image retrieval at the end of the early years, *IEEE Trans. Pattern. Anal. Mach. Intell.*, Vol. 22, No. 12, pp. 1349–1380 (online), DOI: 10.1109/34.895972 (2000).

[61] Smolik, F. and Kriz, A.: The power of imageability: How the acquisition of inflected forms is facilitated in highly imageable verbs and nouns in Czech children, *J. First. Lang.*, Vol. 35, No. 6, pp. 446–465 (online), DOI: 10.1177/0142723715609228 (2015).

[62] Tacca, M. C.: Commonalities between perception and cognition, *Front. Psychol.*, Vol. 2, p. 358 (online), DOI: 10.3389/fpsyg.2011.00358 (2011).

[63] Tanaka, S., Jatowt, A., Kato, M. P. and Tanaka, K.: Estimating content concreteness for finding comprehensible documents, *Proc. 6th ACM Int. Conf. on Web Search and Data Mining*, pp. 475–484 (online), DOI: 10.1145/2433396.2433455 (2013).

[64] Tang, J., Shu, X., Li, Z., Jiang, Y. and Tian, Q.: Social anchor-unit graph regularized tensor completion for large-scale image retagging, *IEEE Trans. Pattern. Anal. Mach. Intell.*, Vol. 41, No. 8, pp. 2027–2034 (online), DOI: 10.1109/TPAMI.2019.2906603 (2019).

[65] Tang, J., Shu, X., Li, Z., Qi, G.-J. and Wang, J.: Generalized deep transfer networks for knowledge propagation in heterogeneous domains, *ACM Trans. Multimed. Comput. Commun. Appl.*, Vol. 12(68), pp. 68:1–68:22 (online), DOI: 10.1145/2998574 (2016).

[66] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D. and Li, L.-J.: YFCC100M: The new data in multimedia research, *Comm. ACM*, Vol. 59, No. 2, pp. 64–73 (online), DOI: 10.1145/2812802 (2016).

[67] Thurstone, L. L.: The method of paired comparisons for social values., *J. Abnorm. Psychol.*, Vol. 21, No. 4, pp. 384–400 (1927).

[68] van Leuken, R. H., Garcia, L., Olivares, X. and van Zwol, R.: Visual diversification of image search results, *Proc. 18th Int. Conf. on World Wide Web*, pp. 341–350 (online), DOI: 10.1145/1526709.1526756 (2009).

[69] Vempala, A. and Preoţiuc-Pietro, D.: Categorizing and inferring the relationship between the text and image of Twitter posts, *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2830–2840 (online), DOI: 10.18653/v1/P19-1272 (2019).

[70] Wang, Y., Stash, N., Aroyo, L., Hollink, L. and Schreiber, G.: Semantic relations in content-based recommender systems, *Proc. 5th Int. Conf. on Knowledge Capture*, pp. 209–210 (online), DOI: 10.1145/1597735.1597786 (2009).

[71] Yanai, K. and Barnard, K.: Image region entropy: A measure of "visualness" of Web images associated with one concept, *Proc. 13th ACM Multimedia Conf.*, pp. 419–422 (online), DOI: 10.1145/1101149.1101241 (2005).

[72] Yee, L. T. S.: Valence, arousal, familiarity, concreteness, and imageability ratings for 292 two-character Chinese nouns in Cantonese speakers in Hong Kong, *PloS one*, Vol. 12, No. 3, p. e0174569 (online), DOI: 10.3389/fpsyg.2016.01907 (2017).

[73] Zhang, M., Hwa, R. and Kovashka, A.: Equal but not the same: Understanding the implicit relationship between persuasive images and text, *Proc. British Machine Vision Conf. 2018*, No. 8 (2018).

[74] Zhao, R. and Grosky, W. I.: Bridging the Semantic Gap in Image Retrieval, *Distributed Multimedia Databases: Techniques and Applications*, Idea Group Publishing, Hershey, USA, pp. 14–36 (2002).