

**Quantifying the mental image
of visual concepts**

Marc Aurel Kastner

Abstract

The semantic gap is defined as the lack of coincidence between the information one can extract from data and the interpretation of that same data. It is a yet unsolved issue for content retrieval and multimedia applications, usually describing issues regarding word choice problems and selecting correct retrieval results, and so on. For example, in applications like image tagging, image captioning, or machine translations, it is often challenging to select the best fitting wording out of a group of candidates.

To create a measurement for perceived differences between concepts and such to quantify the semantic gap of different candidates for word choice problems, this thesis proposes the idea of measuring the visual variety of concepts referring to image data. Abstract or vague input words which have a broad mental image due to being less visually defined would result in a broad feature space, while concrete or visually defined input words result in a rather narrow visual feature space. A system is created which regresses a perceived visual variety score for an input word using visual data analysis. The resulting score describes the input word in its visual variety, approximating the perceived abstractness of that word as a number. For this, two approaches are proposed: Firstly, looking at the relative differences of closely-related words, and secondly as an absolute measurement on a dictionary-level comparison of words.

The first research topic presented in this thesis analyzes the relative visual variety differences of related concepts in a narrow domain by means of a data-driven approach. In this research, existing datasets are reconfigured to create imagesets which reflect the image variety of the real-world. Using the hierarchical relationship of concepts, imagesets for sub-ordinate concepts are aggregated and combined to create imagesets for their composite concepts. As a ratio, a popularity index based on content retrieval engines is used to determine the ratio of sub-concept images. Employing a clustering method on the resulting corpora, the visual feature is quantified to determine a visual variety score for each concept. A crowd-sourced survey is used to

decide ground-truth scores for an expected visual variety for different closely-related concepts. Datasets using different popularity methods are compared to baseline corpora to evaluate the performance of the proposed method.

The second research topic presented in this thesis estimates the absolute visual variety by comparing the variety of visual characteristics across imagesets using an algorithm-driven approach. Using this information, imageability scores for arbitrary words on a dictionary-level are estimated by means of a machine learning model. Thus, in this research, the core assumption of using visual image data for human mental image prediction is applied for the concept of imageability. Imageability is a concept originating from Psycholinguistics, which aims to provide word ratings on a Lickert scale from unimageable to imageable. A large image corpus crawled from Social Media services is analyzed using a mixture of six low- and high-level visual characteristics. Using the cross-similarity across all visual features, a model is trained to regress an imageability score from an input imageset. The corpus is evaluated using imageability dictionaries from Psycholinguistics as a ground-truth. The evaluations compare the proposed method to existing methods using textual analysis instead of image analysis.

As part of the appendix, two dataset visualization projects are outlined, each loosely connected to one of the two research topics introduced above. In these projects, visual datasets originating from either research topic are compared and analyzed regarding their visual characteristics. These projects complement the ideas from the research topics, looking into future directions and applications of the proposed ideas.

In summary, this thesis presents methods to analyze the mental image of words, targeting a way to quantify the semantic gap between vision and language. Chapter 1 gives an overview of the background of this research from various angles. Chapter 2 reviews existing work in the discussed fields thoroughly, giving a comprehensive analysis of the state-of-the-art in this field. The proposed methods for Research Topics 1 and 2 are discussed in Chapter 3 and Chapter 4, respectively. Chapter 5 compares the results of both approaches, outlining the upsides and downsides of each

method for different applications. Lastly, Chapter 6 concludes this thesis by summarizing the research contributions and results found through my doctoral studies.

Acknowledgments

This dissertation is formally submitted for fulfilling partial requirements for the degree of Doctor of Informatics from Nagoya University. This work would not have been possible without the help of many people to whom I owe the sincerest gratitude.

I would first like to thank Prof. Dr. Hiroshi Murase for accepting me into his laboratory twice, beginning with the NUPACE program in 2013–2014 as an exchange student, and then starting 2016 for my doctoral studies. I would like to thank him for his kind help and support throughout my time in Japan.

I would like to express my gratitude to Prof. Dr. Ichiro Ide for his supervision and assistance in this research. His input and insight provided the direction for this research and I am extremely grateful for his contributions and support. He also provided me with the chance to do joint-research with other students resulting in me assisting various research projects around the laboratory. I am grateful for the experiences and knowledge I gained through these opportunities.

I would also like to thank both Assoc. Prof. Dr. Takatsugu Hirayama and Assoc. Prof. Dr. Daisuke Deguchi, who gave invaluable feedback and assistance for various papers and submissions I wrote while studying at Murase Laboratory. Their feedback, often coming from an entirely different viewpoint than my other supervisors, provided an excellent and helpful addition to improve my work.

I thank Prof. Dr. Koichi Takeda for his feedback and insightful comments during the review process of this doctoral thesis.

Last but not least, I would like to thank Dr. Yasutomo Kawanishi for his thorough advice and feedback throughout meetings and reviewing my papers.

Special thanks go to all members of the Murase Laboratory, past and present, who have helped me throughout this journey. The secretaries, Mrs. Hiromi Tanaka and Mrs. Fumiyo Kaba have always been a great help when organizing business trips and

official procedures. I am grateful to Prof. Dr. Shin'ichi Satoh at the National Institute of Informatics, Dr. Frank Nack at University of Amsterdam, and Dr. Kazuaki Nakamura at Osaka University for having meaningful meetings and discussions contributing to my research. I am also grateful to Dr. David R. Wong for discussing research and future career plans. I would like to thank Mr. Kazuki Umemura and Mr. Chihaya Matsuhira for contributing to joint research projects. I would like to thank all of the students, of whom there are too many to name one by one, for their friendship.

I thank the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the German Academic Exchange Division (Deutscher Akademischer Austausch Dienst, DAAD) that have supported me in the form of a Monbukagakusho/DAAD joint scholarship throughout my study as a research student (2016–2017) and Ph.D. student (2017–2020.)

Finally, I would like to thank my family. Despite living far away, my mother has always provided support and encouragement for my education plans and career goals. I would also especially like to thank my partner Ayaka for giving me strength and happiness. Without her, my life would not be the same.

Contents

Abstract	iii
Acknowledgments	vii
Contents	ix
List of Figures	xiii
List of Tables	xv
Abbreviations	xvii
1 Introduction	1
1.1 Motivation: Aim of this research	2
1.2 Background: Vision and language	5
1.2.1 Semantic gap problems	5
1.2.2 Perception of language	6
1.3 Research overview	9
1.3.1 Research topic 1: Relative visual variety differences for concepts in a narrow domain	12
1.3.2 Research topic 2: Absolute visual variety estimation for arbitrary concepts	15
1.4 Thesis structure	18
2 Related Research	21
2.1 Semantic gap	22
2.2 Psychology and human perception	23
2.2.1 Psycholinguistics and perception of words	23
2.3 Multimodal modeling	25
2.3.1 Ontologies	26
2.3.2 Text processing and NLP	26
2.3.3 Tasks quantifying human perception	27

2.4	Applications	28
2.4.1	Use-cases of word ratings	28
2.4.2	Explainable AI	29
2.4.3	Sentiment	30
3	Relative visual variety differences for concepts in a narrow domain	31
3.1	Motivation	33
3.2	Contributions	35
3.2.1	Concept: Visual variety as a way to quantize the mental image of a visual concept	35
3.2.2	Method: Image corpus recomposition to adjust bias of existing image datasets	35
3.2.3	Survey: Establishing ground-truth visual variety labels	36
3.3	Visual variety measurements	37
3.4	Image corpus construction	40
3.4.1	Imbalance of WordNet	40
3.4.2	Recomposition into a balanced corpus	41
3.4.3	Expanding the volume of the corpus	42
3.5	Obtaining the ground truth	46
3.5.1	Crowd-sourced survey	46
3.5.2	Results	49
3.6	Experiment	51
3.6.1	Image corpus creation	52
3.6.2	Survey results	53
3.6.3	Measurement results	54
3.6.4	Rank comparison	55
3.7	Discussion	57
3.7.1	Different popularity metrics	57
3.7.2	Difficulties in corpus construction	58
3.7.3	Ground-truth results	59
3.8	Summary	61
4	Absolute visual variety estimation for arbitrary concepts	63
4.1	Motivation	65
4.2	Contributions	68
4.2.1	Concept: Visual variety for the estimation of imageability	68
4.2.2	Concept: Mixture of low- and high-level visual features to complement semantic information	69
4.3	Imageability estimation	70
4.3.1	Approach	70
4.3.2	Feature selection	72
4.3.2.1	Low-level features	75

4.3.2.2	High-level features	76
4.4	Dataset construction	78
4.4.1	Imageability dictionary	78
4.4.2	Imagesets	79
4.5	Experiment	82
4.5.1	Feature selection	82
4.5.2	Dataset	84
4.5.3	Regression model	86
4.5.4	Evaluation metrics	87
4.5.5	Results	87
4.6	Discussion	94
4.6.1	Performance and feature selection	94
4.6.2	Comparison to text-based methods	96
4.6.3	Dataset	97
4.7	Summary	100
5	Analysis on relative and absolute approaches to visual variety	101
5.1	Core assumptions	102
5.1.1	Personalization	102
5.1.2	Dataset bias	103
5.2	Relative measurement vs. absolute measurement	105
5.2.1	Applications for relative measurements	105
5.2.2	Applications for dictionary-level absolute measurements	106
5.3	Data-driven vs. algorithm-driven approaches	108
5.3.1	Problems of a data-driven approach	108
5.3.2	Problems of an algorithm-driven approach	109
5.4	Reproducibility of published work	110
6	Conclusion	113
6.1	Summary	113
6.2	Remaining challenges and future directions	117
6.3	Closing remarks	120
A	Dataset visualizations	121
A.1	Visualizing Bag-of-Visual-Words models across related concepts	122
A.1.1	Related work	123
A.1.2	Approach	124
A.1.2.1	Dataset	124
A.1.2.2	Visual representations	125
A.1.2.3	Visualization	125
A.1.3	Visualization tool	127

A.1.4	Comparing image regions	127
A.1.5	Summary	129
A.2	Visualization of image sentiment datasets using psycholinguistic ground-ings	130
A.2.1	Approach	130
A.2.1.1	MVSO dataset	131
A.2.1.2	Per-image psycholinguistic scores	131
A.2.2	Visualization	133
A.2.3	Summary	134
	Bibliography	137
	Publication list	149

List of Figures

1.1	Captions which are technically correct might not resemble the mental image of a user	2
1.2	Comparing the concepts of visual distance and visual variety.	3
1.3	Different patterns of retrieval in the same domain	6
1.4	Overview on the field of Psycholinguistics	7
1.5	Core idea of the methodology proposed in this thesis	10
1.6	Example of relative visual variety.	13
1.7	Outline of the relative visual variety estimation process	14
1.8	Example of imageability.	15
1.9	Outline of the imageability estimation process	17
1.10	Thesis structure	19
3.1	A simplified excerpt of the concept tree for self-propelled vehicle from WordNet	32
3.2	Creating a balanced imageset for car based on its subordinate concepts.	34
3.3	Clustering the visual feature space of a concept (e.g., vehicle)	39
3.4	Recomposition of the imageset for a synset.	43
3.5	Flowchart of image corpus construction.	44
3.6	User interface for the crowd-sourced survey.	47
3.7	Examples for different corpus recompositions of the same synset.	51
3.8	Visualizing the overall trend of each corpus.	54
3.9	Stability of Spearman Rank correlation results for different ground-truths.	60
4.1	Core concept of word imageability.	65
4.2	Flowchart of the imageability estimation process.	73
4.3	Flowchart of dataset acquisition of visual data for words and their imageability labels.	81
4.4	Scatter plot of predicted imageability scores.	85
4.5	Example of image datasets and their predicted imageability scores.	86
A.1	Example of a visualized synset.	123
A.2	Process of selecting common visual words.	126

A.3	Example of the common keypoint visualization for the synset “truck”.127	
A.4	Visualization tool shows detailed information as well as the key- points when hovering the mouse over the datapoints.	128
A.5	Process of calculating per-image psycholinguistic scores.	132
A.6	Main user interface of the proposed visualization.	133
A.7	Spatial embedding can be colored in different ways based on their calculated individual scores or dataset annotations.	135

List of Tables

3.1	Examples of estimated visual variety results.	55
3.2	Quantitative analysis of the proposed method.	55
3.3	Comparison of different Web popularity measurements.	58
4.1	Qualitative analysis for different sets of visual features.	88
4.2	Qualitative analysis for different training dataset sizes.	90
4.3	Quantitative results for different regressors.	90
4.4	Feature comparison for abstract words vs. concrete words.	91
4.5	Feature comparison for different parts-of-speech.	92
4.6	Imageability prediction results of the proposed method.	93

Abbreviations

2D	2 Dimensional
3D	3 Dimensional
AMT	Amazon Mechanical Turk
ANP	Adjective-Noun Pair
API	Application Programming Interface
BoVW	Bag of Visual Words
COCA	Corpus Of Contemporary American English
CPU	Central Processing Unit
CV	Computer Vision
GIS	Google Image Search
GPU	Graphics Processing Unit
GT	Ground Truth
GTS	Google Text Search
HSV	Hue-Saturation-Value (Color space)
LIWC	Linguistic Inquiry and Word Count
MAE	Mean Absolute Error
MRC	Medical Research Council
MSE	Mean Square Error
MVSO	Multi-lingual Visual Sentiment Ontology
NLTK	Natural Language ToolKit
NLP	Natural Language Processing
RBF	Radial Basis Function
SE	Sketch Engine

SURF	S pedeed Up R obust F eatures
SVM	S upport V ector M achine
UMAP	U niform M anifold A pproximation and P rojection for Dimension Reduction
XAI	E Xplainable A rtificial I ntelligence
YFCC	Y ahoo F lickr C reative C ommons
YOLO	Y ou O nly L ook O nce (Name of pre-trained model)

For my family

Chapter 1

Introduction

In recent years, the growth of multimodal 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, however, is vague, and the semantics of tagging might change depending on the choice of words. A rather abstract tag like “vehicle” might not describe an image of a specific motorbike type particularly well. The model name of the said motorbike, in contrast, might be too specific, as an average user might not have a mental image of it. 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 this gap, it is essential to have a deep understanding of how vocabulary and their visual representations connect. As a first step towards such an understanding between vision and language, this thesis discusses the visual analysis of Web-crawled image data to quantify this perceived variety of different visual concepts.

In this chapter, the purpose and motivation of this doctoral research are first outlined in Section 1.1. Section 1.2 discusses the history of research related to semantic gap problems in vision and language. Next, Section 1.3 briefly summarizes each research topic discussed in this dissertation. Lastly, Section 1.4 gives an overview of the structure of this thesis.



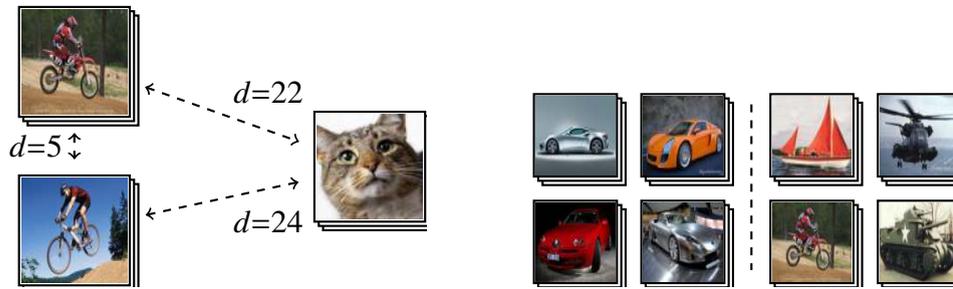
Caption: *This is a red vehicle.*

Figure 1.1: Captions which are technically correct might not resemble the mental image of a user.

1.1 Motivation: Aim of this research

In recent Web developments, whether it is media retrieval, automated generation of content, or user-interaction through applications like Social Media, virtually any application uses a combination of multiple modalities. Typical types of modalities in multimedia research are text, image, video, and audio, among others. When developing such applications, different media need to be connected. This comes with a need for an understanding on how text, image, video, and so on, connect.

Multimedia modeling is an approach to harmonize different media and systems. A semantic understanding of how different media interact is needed to generate natural results. Furthermore, as the user of most systems is a human, the perception of how vision and language interact is another dimension to keep in mind. Image captioning, for example, is the application of automatically generating a text-description for a given image. Considering Fig. 1.1, the caption “*This is a red vehicle*” might be technically correct, but the mental image of a red vehicle might be closer attached to a red car or truck. It is also not helpful in understanding the situation the vehicle is in, nor does it provide any additional knowledge about the situation of the image, the driver, its characteristics, and so on. As such, a caption more closely resembling the image of a user might rather be “*A red helicopter*”, or even more concrete “*A red coast guard helicopter in the sky*”. Following, a better understanding about how a human perceives different information, and which words thus seem abstract, specific,



(a) Visual distance: How different are two concepts? (b) Visual variety: In how many ways can a concept be visualized?

Figure 1.2: Comparing the concepts of visual distance and visual variety.

helpful, verbose, and so on, would be valuable information to provide the system for better word choices.

This thesis looks into this problem by quantizing the perception of humans regarding vision and language. To illustrate this idea, let us first jump back and discuss a common concept for the comparison of two things: Distance. For the visual distance, for example, visual features are compared to calculate a distance between two images. Similarly, one could aggregate this for a whole dataset to find the distance between two visual concepts. Such a measurement is calculated *between* two concepts, so, e.g., a mountain bike would be closer to a racing bike than to a cat (Fig. 1.2(a)).

Meanwhile, when thinking about an individual concept in a group of others, there is, however, no metric to describe its characteristics in the big picture of things. One could ask the questions: Is one concept *broader* than the others? Is it more *visually diverse* than other concepts? These questions are interesting on their own and could be of benefit for applications like the visual diversification [21] in image retrieval. There is also a relationship to how humans perceive these concepts differently. A *visually diverse* concept might tend to be less clearly defined and thus more abstract, while a *visually distinct* concept might be easier to grasp and more concrete. A word like `vehicle` is more visually diverse than a more concrete term like `car` (Fig. 1.2(b)). For applications having to choose between different output candidates,

like word choice problems, there is often no real answer, which candidate to prefer. Rather, it depends on the application —as a more vague word choice might be better for some, while a very specific one might be better for others. This leads to the research of *visual variety* measurements, targeting the quantification of such a metric. This thesis aims to propose a perception-based scale, describing a concept in the big picture of other concepts.

This problem is related to semantic gap issues in vision and language. Following, the next section will introduce the background on this by first defining semantic gap problems in multimedia applications and then further discussing the perception of language.

1.2 Background: Vision and language

This section discusses the background of this research regarding the semantic gap between vision and language. In this thesis, a *mental image* is defined as the “visual experience where the content does not directly relate to any afferent stimulus but is derived from (working) memory” [96]. It is related to *visual perception*, which is the “visual experience where the content reflects and is caused by an afferent physical stimulus” [96]. As such, the mental image is a mentally visualized experience, while the visual perception is caused by an actual physical stimulus (typically through the eyes.) Mental imagery as a trait is one of the primary human mental events that allows remembering, planning in the future, navigate, and making decisions [98].

This thesis deals with the estimation of a score that describes the span of the mental image regarding a concept. Furthermore, the target is not the estimation of that of an individual user, but rather the average across society.

1.2.1 Semantic gap problems

The semantic gap is “the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation” [83]. In layman’s terms, it means that depending on the application, users are interested in vastly different outcomes. This has implications for most multimodal applications, like any media retrieval, tagging, and description applications, among others.

Let’s consider the example of image retrieval, where the user searches different queries including *car*. For the query “*car*”, pretty much any typical image of a car will suffice and is appreciated as a result. If the query is changed to “*car with three wheels*”, however, only a certain specific sub-class is wanted. If the query is instead changed to “*car like Lamborghini Aventador*”, only a very specific car is filtered. While all queries search for a *car*, the actual retrieval can be narrow category-search,



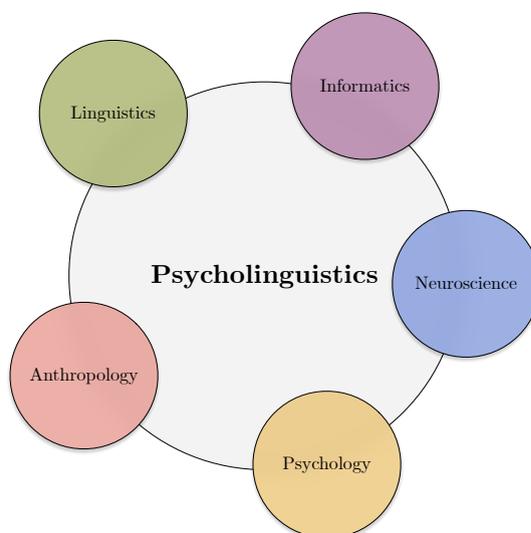
Figure 1.3: Different patterns of retrieval in the same domain.

wide category-search, or even a target-search. This idea is illustrated in Figure 1.3. From a single word, like *car*, it is thus impossible to decide which outcome would be appropriate. Rather, one needs to look at the surrounding context to decide which selection of images would be the best fitting.

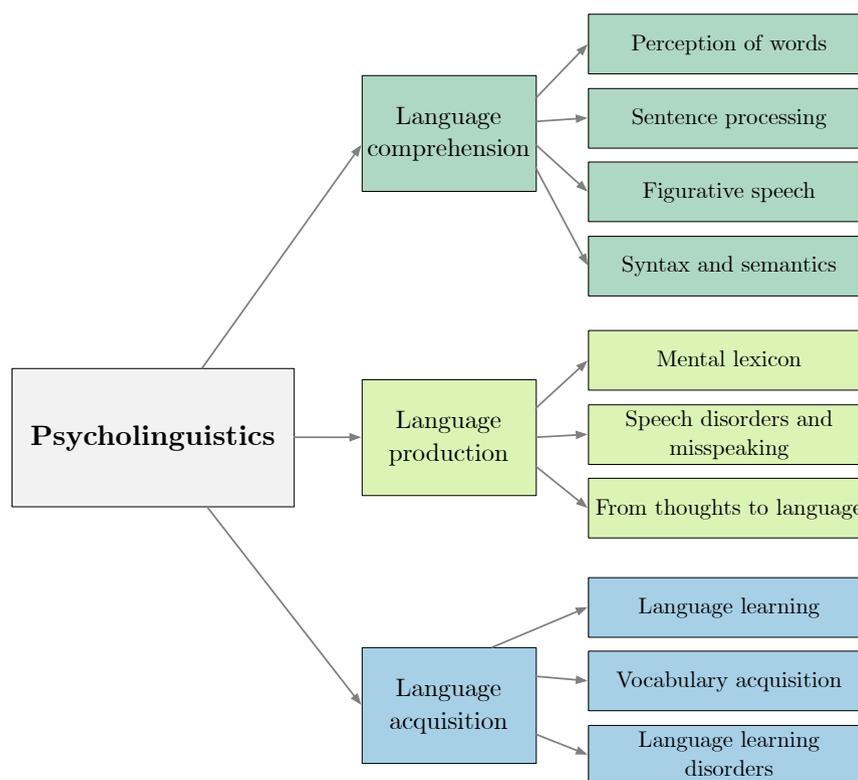
This example also works in the opposite direction. When considering an example like the tagging or captioning of images, which of the example queries would be the best fit? Depending on the applications, a tag like *car* might be just fitting, too abstract, or too concrete. Which one it is; that needs to be decided by the use-case. However, there are few metrics to decide whether a certain candidate might be concrete or abstract in the big picture of things.

1.2.2 Perception of language

The field of Psycholinguistics is a branch of Cognitive Sciences. As a cross-disciplinary area, it researches the connection between neuroscience, computer science, linguistics, anthropology, and psychology. The target of research commonly is the understanding of three areas: language production, language comprehension, and language acquisition from a Psychological point of view. An overview of the field is shown in Figure 1.4.



(a) Positioning the cross-diciplinary field of Psycholinguistics.



(b) Directions of research in Psycholinguistics (Excerpt of fields relevant to this thesis.)

Figure 1.4: Overview on the field of Psycholinguistics. (Based of lecture notes PSY301 at New Mexico State University. [2])

As a study of Psychology around language, the research in Psycholinguistics also looks at the perception of language and words. For example, the learning process of language is analyzed for both first- and second-language learners, the latter looking into how already known languages influence the learning processes of new languages. In other research, Psycholinguistics looks at language understanding during conversation or reading. With eye-tracking and similar sensors, the influence of grammatical errors or unknown words during language understanding can be measured. Both these researches get additional twists if looking at either children or adults with language learning disabilities.

In the 1960s to 70s [28][78], there was the first research towards the concepts of concreteness and imageability in the English language. In these concepts, words are rated on a Lickert scale; Test subjects are asked to judge the concreteness and imageability of different words from 1 (very low) to 7 (very high). As a result, dictionaries have been created, listing such ratings for a selection of words in various languages [30][58][59][60].

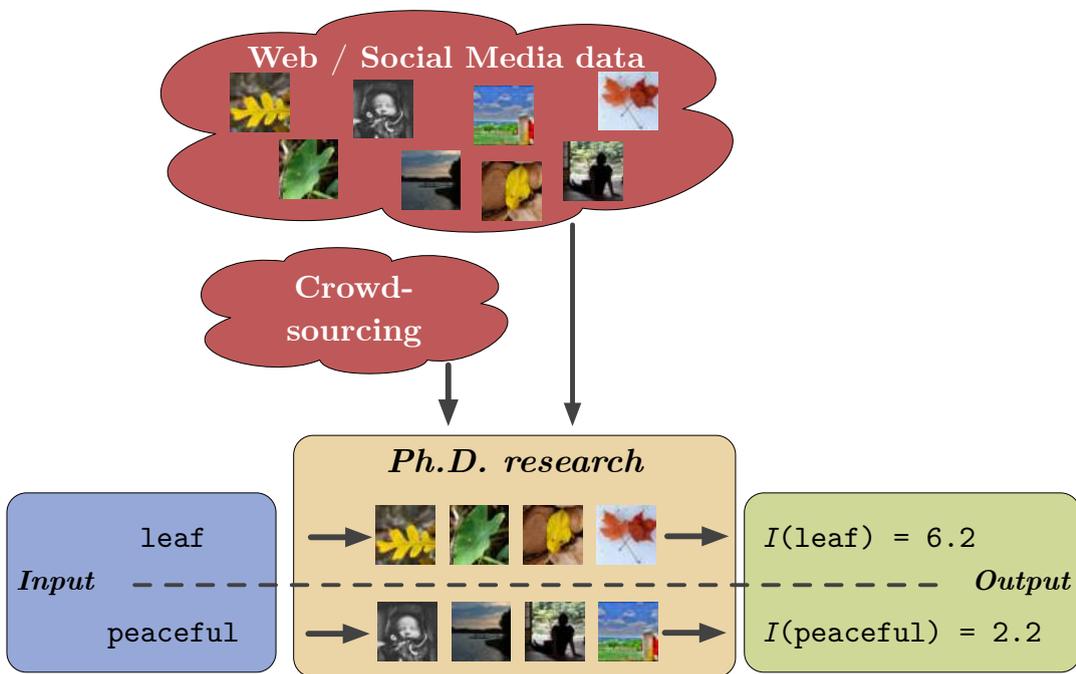
Among these concepts, *imageability* is defined as “The ease with which a word gives rise to a sensory mental image” [28]. Following, there is a relationship between the mental image of a word and its imageability scoring. Research [99] also concludes that imageability and concreteness should be distinguished, both experimentally and theoretically. An example of a term with a high concreteness and a low imageability rating is *astrolabe*: People know that it describes a concrete object, but have no mental image of it [97].

1.3 Research overview

The semantic gap is an essential yet unsolved issue for content retrieval and multimedia applications. When it comes to word choice problems such as image tagging, image captioning, or machine translations, the perceived abstractness of concepts can be an indicator of which word to choose out of a selection of synonyms, for example. The main topic of this thesis is to measure the visual variety of different concepts as an indication of the perceived mental image of the said concept. A simple example of such a measurement could be stated as follows: If having multiple words such as `vehicle`, `car`, or `sports car`, what is the visual semantic gap between them in terms of how they are perceived by a human? As such, this idea in the field of multimedia modeling could be used as an evaluation metric for created tags or captions, or as semantic information between vision and language for use in other research and applications.

The core idea of this research is to quantify the perceived mental image of input words or concepts. Abstract or vague terms often have a broader mental image due to them being less visually defined. Similarly, such abstract or vague terms (e.g., `algebra` or `peaceful`) would have broader visual characteristics — simply because they are also less explicitly defined. In contrast, concrete or visually well-defined input terms (e.g., `leaf` or `car`) result in a rather narrow visual characteristics, mostly because most related images depict the same object.

This idea is illustrated in Figure 1.5. Visual data retrieved from Web and Social Media is analyzed regarding the variety of its visual features. With this, a system is created which regresses a perceived visual variety score for an input word (Fig. 1.5(a)). The output represents the input word in its visual variety, approximating the perceived abstractness of that word, as a numerical score or ranking. The proposed idea 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 (Fig. 1.5(b)). As such, if gathering a sufficient number of images related to a word or concept, the visual feature space will converge towards the average mental image



(a) Goal of the research topic introduced in this thesis.

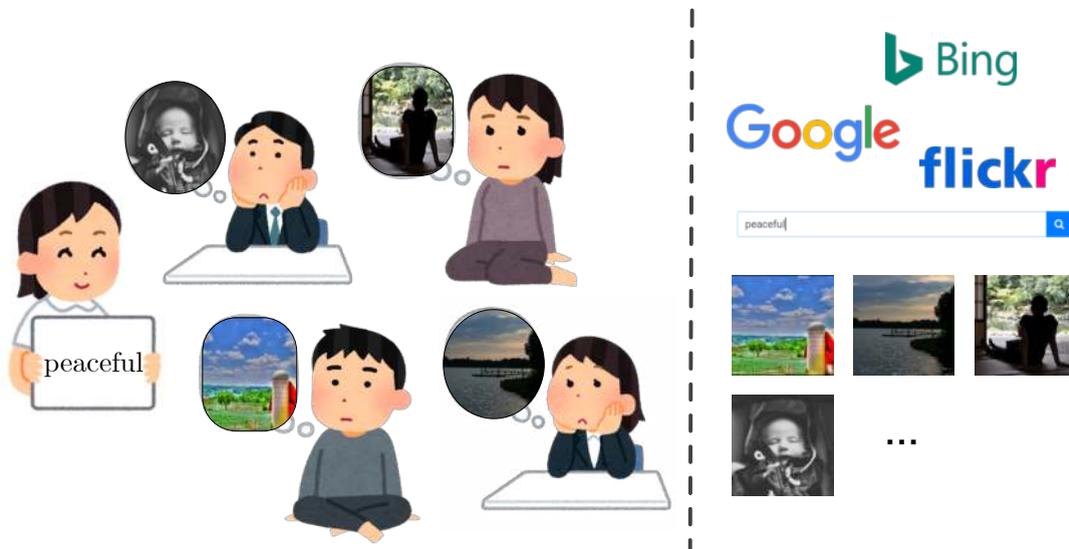
(b) Mental image of words across society vs. images from the Web. As a core assumption, the mental image of a term such as *peaceful* as perceived by a variety of people can be loosely approximated by crawling Web-based services for images regarding the said term.

Figure 1.5: Core ideas of the methodology proposed in this thesis.

of the said word or concept. In order to tackle the task of mental image quantification, it can be divided into the following two sub-tasks of relative and absolute measurements.

First, as a relative measurement, the goal would be to find granular differences of related concepts. The perceived differences between, e.g., `vehicles`, `cars`, or `tanks` would give an indication on which to use for a target application. For example, when asked about a `vehicle`, more people will presumably think of `cars` than of `tanks`. Following, the concept of `cars` might have more influence on the overall mental image of the term `vehicle` than the concept of `tanks`. This relative understanding of multiple concepts in a narrow domain would benefit word-choice problems in multimedia applications. One target application I have in mind when approaching this research is image tagging. In image tagging applications, there is often a choice between many possible candidates: An image of a car could be annotated with `vehicle`, `car`, or `sports car`, depending on the context or use-case. The relative distance can give an indication on which word could be the most appropriate for a certain use-case. In order to implement this approach, a dataset that actually reflects the ratio of images close to the human's expectation is needed. Since existing datasets are somewhat biased from this point of view, they need to be reconfigured. Section 1.3.1 briefly introduces this idea as the first research topic, using a data-driven approach to analyze the relative visual variety differences of related words.

Second, as an absolute measurement, the goal would be to find a general trend of the perceived size of the mental image. As such, one could discern concepts of rather high perceived variety from concepts with a low perceived variety. For example, the concept `car` is visually clearly defined, but the concept `peaceful` has no such clear image. While a direct comparison is somewhat difficult as the unrelatedness makes finding a common reference problematic, an absolute measurement can help indicate global trends of less related concepts. Using images accumulated from Social Media, an analysis of the visual feature space thus gives knowledge about the perceived variety. Analyzing, e.g., `car` results in a narrow visual feature space, while `peaceful`

becomes a very noisy feature space. As a target application for this research, I am considering multimodal approaches analyzing the relationship of text and images. Having applications like image tagging in mind, a metric on the overall trend of imageability helps in understanding which concepts can be visually depicted. This is useful to improve the quality of auto-generated texts. A comprehensive analysis across low- and high-level visual features can be used to quantify such characteristics for each word. Section 1.3.2 briefly introduces this idea as the second research topic, using an algorithmic approach to compare a variety of visual characteristics across datasets to estimate the absolute imageability of words on a dictionary-level.

The first research topic looks at the semantic gap as a relative measurement in order to get a better understanding of related concepts. The second research topic, in contrast, looks at an absolute measurement in order to find a general trend of the perceived gap even for unrelated concepts. Multimodal applications could furthermore profit from using both types of knowledge; e.g., by first getting an understanding of the overall trend of unrelated concepts with an absolute measurement, before tackling the more fine-granular word choice problems using a relative measurement. As such, solving both these problems would allow analyzing the quantification of the mental image from its diverse angles, solving the proposed aim of this thesis. As both viewpoints are connected to human perception on different scales, this allows for a better and more thorough understanding of text-image relationships in various applications.

1.3.1 Research topic 1: Relative visual variety differences for concepts in a narrow domain

For multimedia applications such as image tagging, a comparison of the abstractness of closely related terms becomes crucial to understand the differences between, e.g., a `vehicle`, a `car`, or a `sports car`. Figure 1.6 shows an example of the relative visual variety for a selection of related concepts. In the example, the words in the

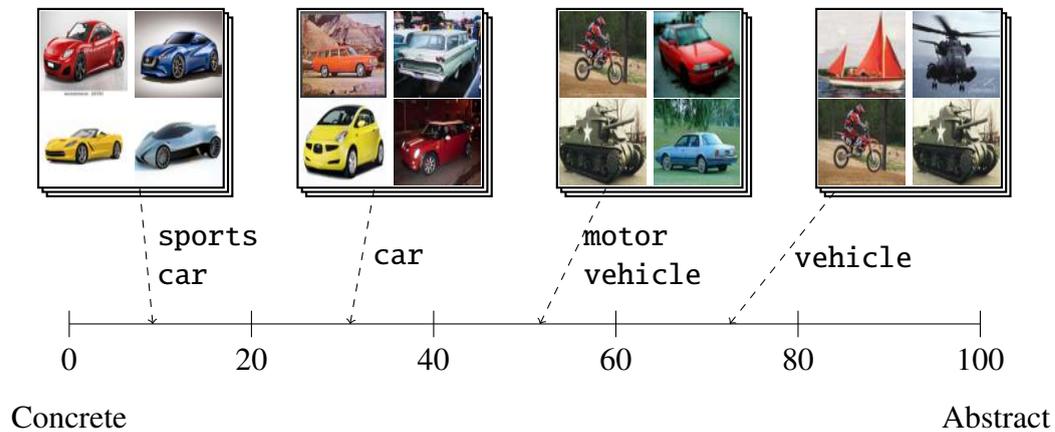


Figure 1.6: Example of relative visual variety.

domain of `vehicles` are evaluated in on a 1-dimensional scale from concrete to abstract. The parent term `vehicle` is used as a reference point to quantify the variety relative to related sub-concepts. With this, this research targets a relative measurement of variety for similar concepts in a limited domain.

In the motivating Figure 1.5(b), the mental image of a concept was compared to what people imagine when thinking about the said concept. Following, the composition of a concepts' imageset would need to correlate with the expected contents. In this research topic, having this in mind, new imagesets are created for each concept. Reorganizing the composition of an existing but biased imageset, the approach creates a recomposed and expanded imageset. Using these *idealistic* and less biased imagesets, the relative visual variety is then computed.

The basic approach is illustrated in Figure 1.7. For each term, a collection of subordinate concepts is collected through WordNet [45]. Each of these subordinate concepts contributes, to a degree, to the overall image of the said term. For each subordinate concept, a large number of images is crawled. Using an Web-based API, a popularity metric is calculated. The ratio of two subordinate concepts' popularity decides which concepts are more common in the average mental image of the said term than the others. With the popularity metric as a basis, a new imageset for each concept is created. Lastly, the visual feature space is extracted and clustered. The number of clusters decides the relative visual variety for each term.

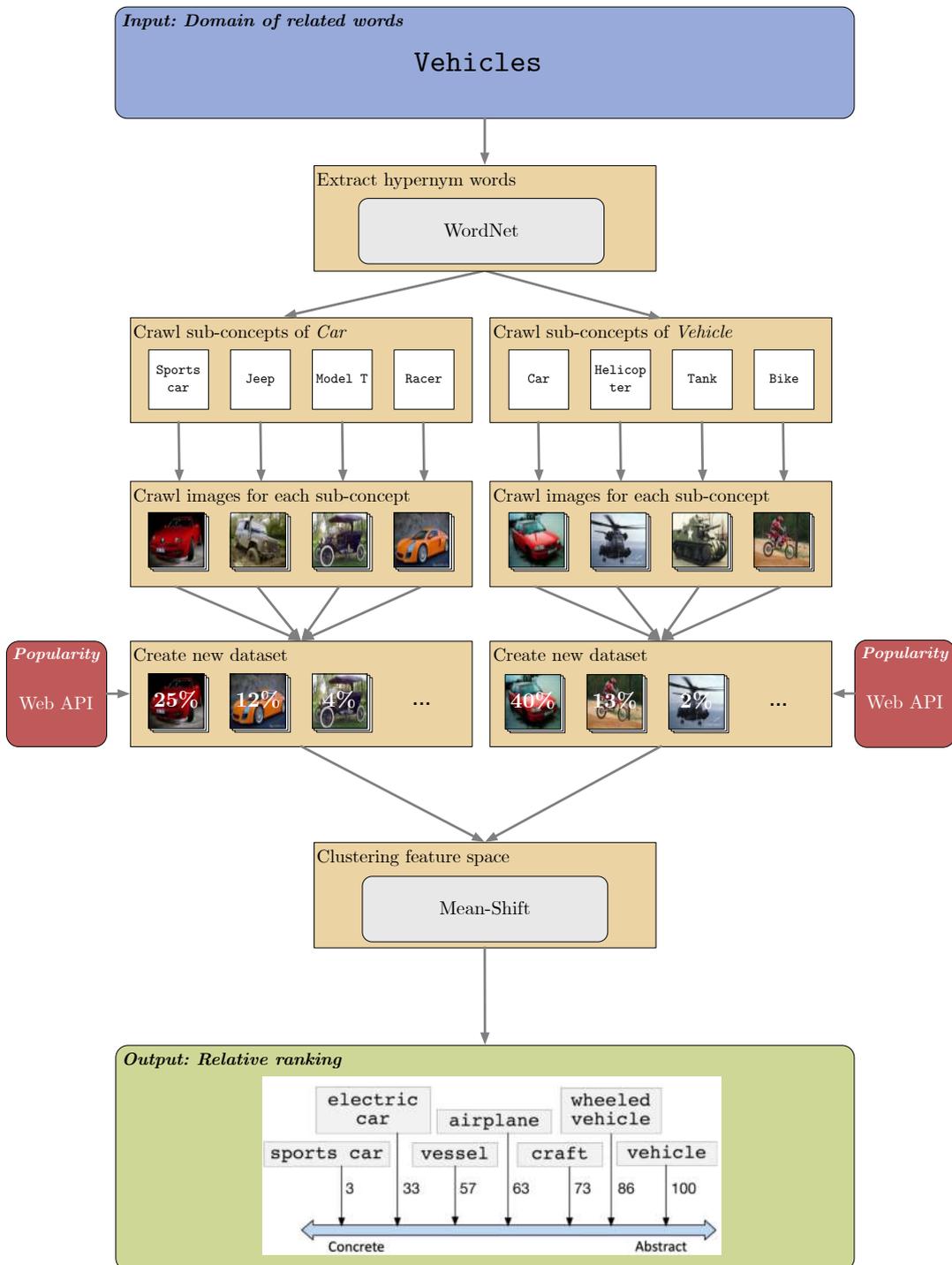


Figure 1.7: Outline of the relative visual variety estimation process.

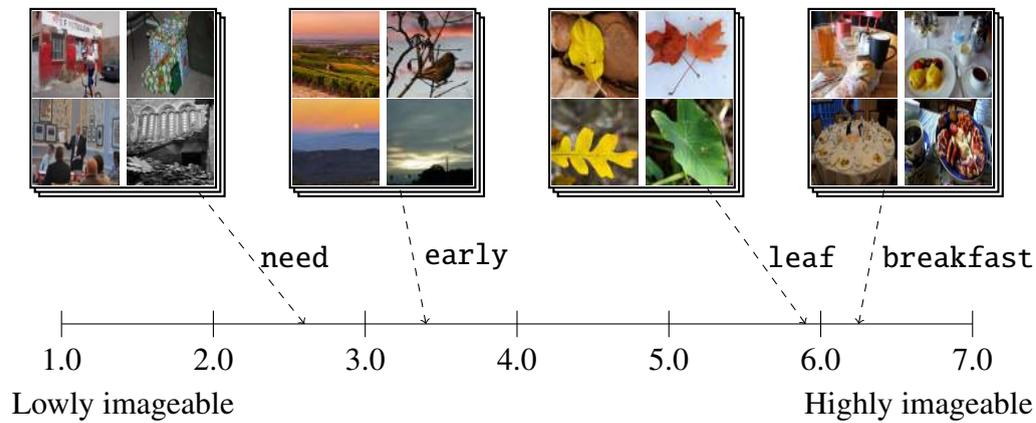


Figure 1.8: Example of imageability.

1.3.2 Research topic 2: Absolute visual variety estimation for arbitrary concepts

In Psycholinguistics, which is a research field that crosses Psychology and Linguistics as illustrated in Fig. 1.4, there are existing dictionaries including word ratings for the English language. One of these metrics named *imageability* is a metric describing whether a word is easy or hard to imagine. On a Lickert scale, e.g., from one to seven, people are asked to judge the imageability of words from highly imageable to lowly imageable. Figure 1.8 shows examples for imageability on a selection of words. This concept is related to the idea of visual variety of datasets as discussed before. In this research, the core ideas of visual variety are extended for this use-case in the field of Psycholinguistics. According to Richardson [99], there is a relationship between a mental image of a word and its imageability scoring, connecting the ideas of the core assumption to the measurement of imageability. By analyzing the overall visual feature space of one dataset per input word, a model is trained to regress an imageability score for each word.

The basic approach is illustrated in Figure 1.9. For each word, a dataset is crawled from Social Media services. Using a large number of images, the visual feature space for each word is extracted from a variety of low- and high-level visual features. The resulting feature histograms are cross-compared to create a similarity matrix which is used to train a model. For ground-truth annotations, existing imageability

dictionaries are used. The proposed system can predict imageability for new words not in the dictionary by simply crawling images of those words from the Web.

A core difference that distinguishes this research from Research topic 1 is the variety of words in a general dictionary. While a relative measurement deals with granular differences between related concepts, and thus allows for a common reference point, a dictionary-level comparison of concepts deals with unrelated words. As such, it does not only need to deal with the difference of `car` and `vehicle`, but also that of `pizza` and `peace`. In the case of imageability, the scale includes any type of word from a general-purpose dictionary, potentially all words of the corresponding language.

As there are existing imageability dictionaries [30] [58] manually created as part of Psycholinguistic research, there is a relatively large number of ground-truth annotations to be used for training and evaluation. The method proposed in this research could be used for automated imageability estimation, which would be useful to further extend existing dictionaries without the need of manual labour. Accurately estimated scores for a larger part of the general dictionary would benefit various research in natural language processing and multimodal applications.

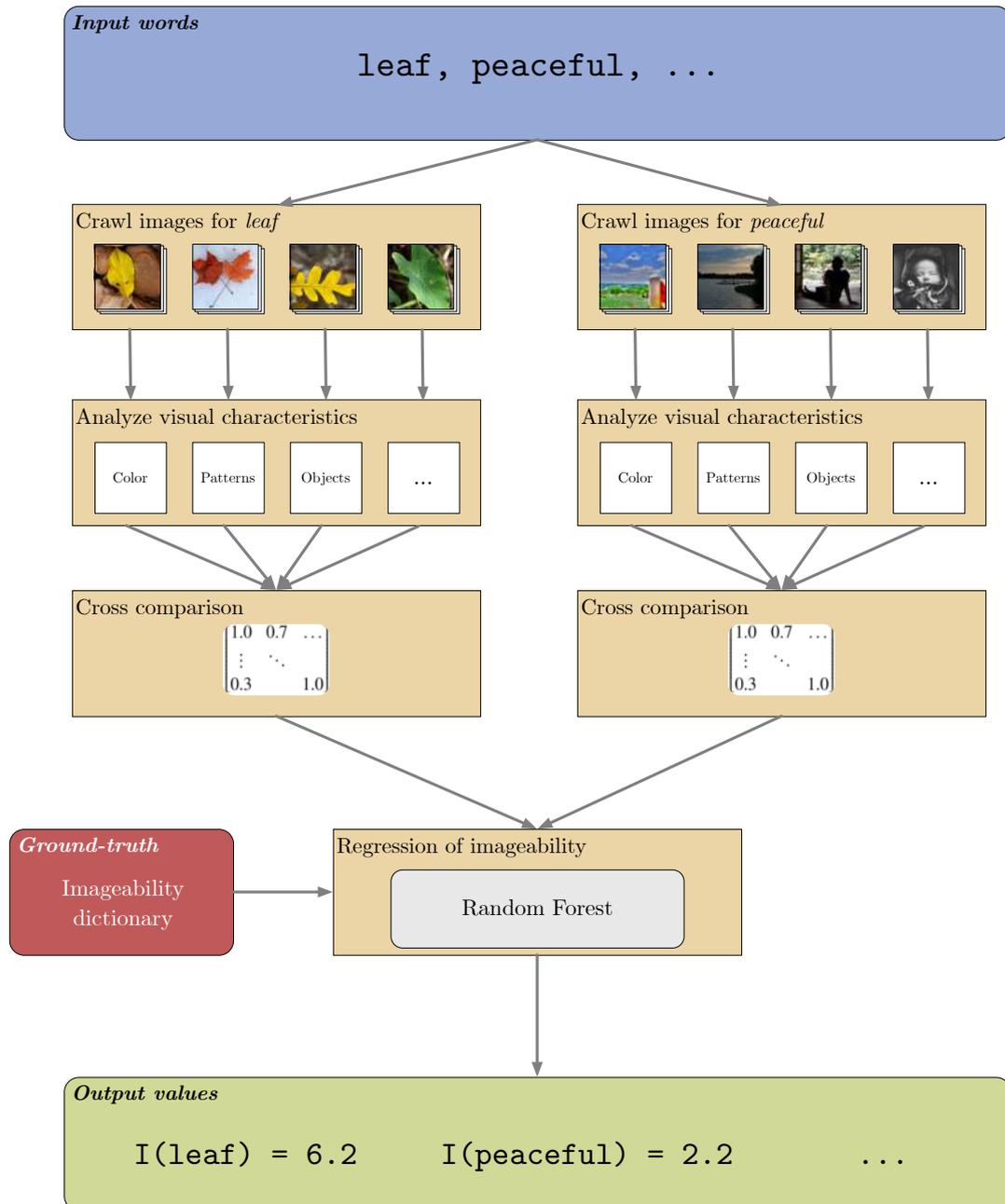


Figure 1.9: Outline of the imageability estimation process.

1.4 Thesis structure

This thesis contains six chapters and one appendix. The relationships between the different chapters of this thesis are visualized in Figure 1.10.

This Chapter 1 discussed the motivation of this doctoral research and gave an overview on the background of research involving vision and language from three angles: Semantic gap issues, Psycholinguistics and human perception of words, and Multimedia applications. Two research topics are proposed to solve issues in these fields by quantizing the visual variety of datasets for word perception understanding. Chapter 2 reviews existing work in the discussed three fields thoroughly, giving a comprehensive analysis of the state-of-the-art on this research. Chapter 3 discusses the first research topic outlined in Section 1.3.1, using a data-driven approach to analyze the relative visual variety differences of related words. Chapter 4 discusses the second research topic outlined in Section 1.3.2, using an algorithmic approach to compare a variety of visual characteristics across datasets to estimate the absolute imageability of words on a dictionary-level. Afterwards, Chapter 5 compares the results of both researches, outlining the upsides and downsides of each proposed method for different applications. Lastly, Chapter 6 concludes this thesis by summarizing the research contributions and results found through these studies. In addition, future research directions, remaining challenges, and applications that can be built from the results will be discussed.

As a supplement to the main part of the thesis, Appendix A introduces two dataset visualization projects for the analysis of datasets used in the main thesis. Both visualizations explore visual and semantic relationships of the datasets used in research topics 1 and 2. Project A.1 looks at visual similarities across datasets created for research topic 1, visualizing variety differences across synset siblings in a hierarchy of words. Project A.2 looks at psycholinguistic features in text associated to Flickr images to find similarly described images and thus candidates for similarly perceived images. It embeds imageability scores, among other features, to create a spatial psycholinguistic space in the dataset used in research topic 2.

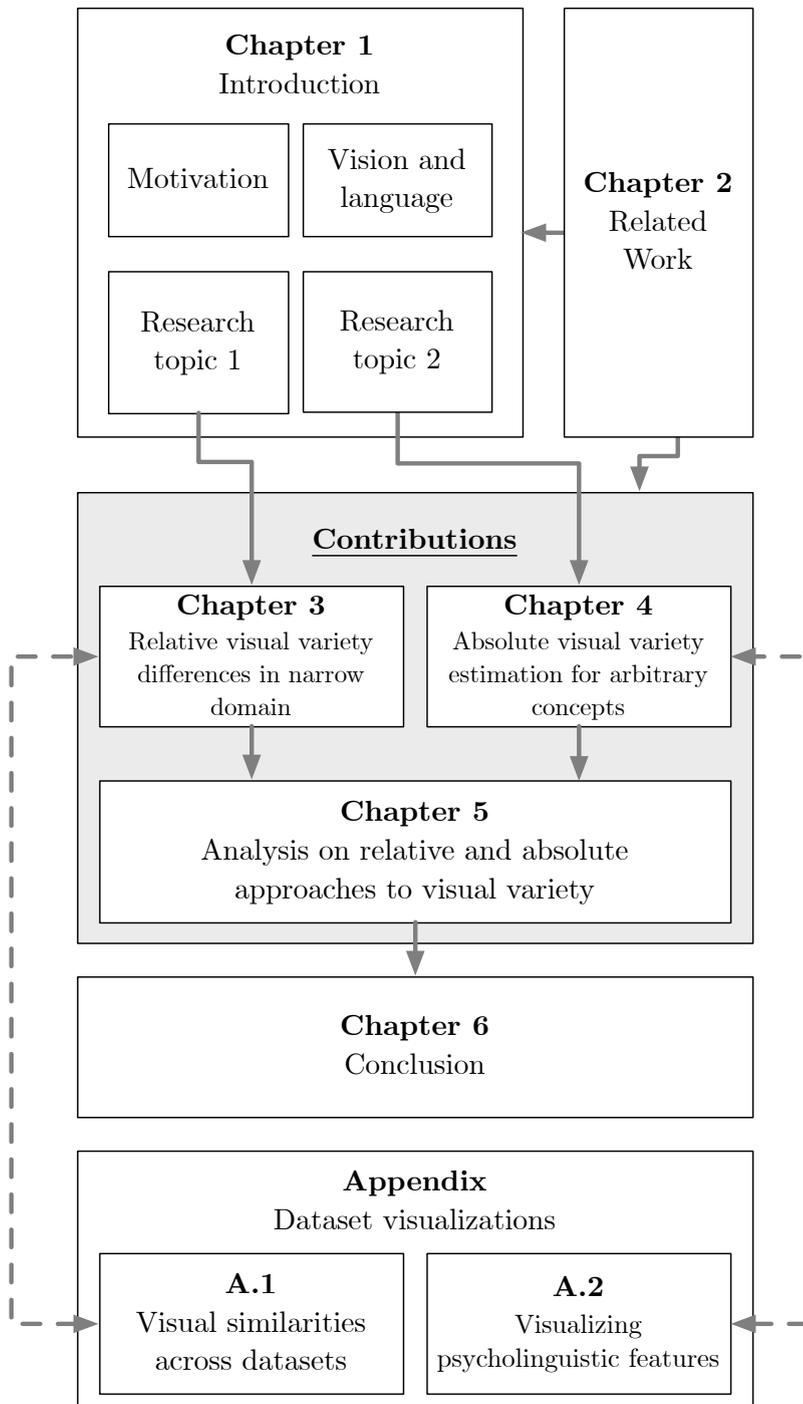


Figure 1.10: Thesis structure.

Chapter 2

Related Research

In Chapter 1, the motivation of this doctoral research has been introduced, discussing two sub-tasks to solve in order to quantify the mental image for use in multimodal applications. Following, this chapter gives an overview of existing literature related to both Research Topics 1 and 2, discussing the semantic gap, or in other words, the understanding of human perception, between vision and language. Section 2.1 will discuss existing research regarding semantic gap problems in general. This includes overviewing work as well as some ideas towards narrowing and solving the semantic and sensory gaps. Section 2.2 discusses related work in the fields of Psychology and Psycholinguistics regarding human perception in general, but also towards language and words. Section 2.3 will discuss multimodal modeling research towards semantic knowledge or semantic embeddings. Lastly, Section 2.4 will discuss a variety of multimodal applications, which are use-cases of such models, and impacted by the proposed research.

2.1 Semantic gap

The *semantic gap* in content-based image retrieval received most attention through the work by Smeulders et al. [83]. Discussing the definitions of semantic gap and sensory gap, this survey paper discussed a variety of usage patterns of image retrieval. Following, Nack [89] described the semantic gap between rich meaning users expect and the shallowness of content descriptions as a crucial obstacle to overcome for future applications. Dorai and Venkatesh [90] cited the “manipulation of affect and meaning”, “the representation, extraction, and synthesis of expressive elements”, and “metrics to assess automatic extraction techniques” as the biggest future challenges to solve the semantic gap in multimedia applications.

Over time, there has been much research in narrowing or bridging the semantic gap for image retrieval and recommendation system purposes. Zhao et al. [87][88] introduced color histograms and color anglograms into Web document retrieval to improve its performance. Cheng et al. [85] introduced semantic visual templates as a means to personalize content-based recommendation systems. They bridged the semantic gap by including a personalized view of concepts into the template. Wang et al. [86] used semantic relations to improve the relatedness of recommendations in general. Jiang and Conrath [35] used corpus statistics and lexical taxonomies like WordNet [45] to determine a semantic distance measurement which helps to understand textual semantic differences. Budanitsky and Hirst [46] continue this work by comparing five different semantic measures obtained through WordNet.

2.2 Psychology and human perception

Perception has been previously 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” [50]. Cahen and Tacca linked the concepts of perception and cognition in their survey paper [95]. They described perception as the input to cognition, discussing how the mechanisms communicate with each other. Tacca [84] further argued, that the cognition, so the process of understanding through thought, actually has an influence on how we see things, and thus influences the perception. Therefore, the communication between both is described as a dialog. Montemayor and Haladjian [94] discussed them as largely independent concepts, but still argue that both affect each other through evolutionary influence.

Regarding visual perception and mental imagery, Dijkstra et al. [96] found that both mechanisms share the same neural structures.

2.2.1 Psycholinguistics and perception of words

In 1968, Paivio et al. [28] first proposed the concepts of *imageability* and *concreteness*, along a third metric of *meaningfulness* as measurements for human perception of natural language. Since then, there has been ongoing research, connecting language understanding and language acquisition to the imageability of words and concepts. While Paivio treated imageability and concreteness fairly similar, Richardson [78] discussed the difference between both concepts. He concluded that imageability and concreteness should be distinguished, both experimentally and theoretically.

Smolik and Kriz [11] discussed implications of the imageability of verbs on grammar usage for different contexts, which could provide helpful knowledge to create more natural language depending on the context. It is considered 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. There is also a relationship on imageability of words to age of acquisition and reading comprehension, especially relevant to children [51][55]. Due to this, Jones [53] discussed further use of imageability in the research of dyslexia. Schwanenflugel [54] discussed the relationship of text difficulty and concreteness, when it comes to abstract words, as it represents the fundamental semantic distinction between them. In Neuropsychology, there was research by Giesbrecht et al. [52] discussing the neurological process of word understanding in relation to their imageability. There are imageability dictionaries for English [30][58] as well as other languages [59][60]. However, the dictionary creation process is labor-intensive, as the annotations are commonly obtained through crowd-sourcing or user studies involving test subjects.

2.3 Multimodal modeling

In Multimedia research, the analysis of visual concepts has been ground for multiple works. Prominently, this research involves estimating or quantifying the relationship of different concepts. Nakamura and Babaguchi [23] measured the distance for visual concepts with an adaptive weighting for multiple visual features. Furthermore, Nagasawa et al. [29] 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. Other work by Yanai and Barnard [27] analyzed image region entropy to identify *visualness* of adjectives, later continued by Kohara and Yanai [41] to analyze Adjective-Noun Pairs (ANPs). Divvala et al. [42] proposed a method to analyze visual features to create visual knowledge databases with unsupervised crawling. Tang et al. [4] looked at social-aware tagging by including user-information into the training to remove noisy and unimportant tags. Furthermore, there is also research in using deep networks to model cross-domain information between text and images [91][92][93].

Van Leuken et al. [21] performed 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 that are very closely related. Next, they select 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.

In the following, I will outline research topics inside Multimedia research related to this thesis, starting with research regarding hierarchical ontologies. Next, research connecting the field of Natural Language Processing (NLP) with multimedia applications is discussed. Finally, I will outline recent popular tasks in Multimedia workshops and benchmarks related to this field.

2.3.1 Ontologies

Kawakubo et al. [18] proposed an idea on how to automatically create an ontology for visual features. They cluster similar images to create a hierarchical structure of related visual concepts. Meanwhile, Inoue and Shinoda [22] tried to analyze the ontological relationship of visual concepts by directly incorporating lexical relationships. They calculate a weighting, which describes how much a hyponym has a visual influence on its hypernyms. Ordonez et al. [10] use an ontology to improve the tags of WordNet based object category prediction by replacing unnatural descriptions like “grampus griseus” with more straightforward *entry level categories* like “dolphin”.

2.3.2 Text processing and NLP

In the field of NLP, some researchers have been working towards the estimation of imageability or concreteness using text data mining techniques. Ljubescic et al. [8] created a word embedding predicting the concreteness and imageability of words within and across languages, evaluating with English and Croatian. Similarly, Charbonnier and Wartena [5] predicted the word concreteness and imagery from image captions using text data-mining methods.

Computational linguistics can also profit from these metrics as a complement to sentiment embeddings. Kiela et al. [19] significantly improved the performance of multimodal embeddings by looking at image dispersion. This suggests that the variety of images in datasets have strong influence on the perception of words and thus their embedding. Hessel et al. [7] used the multimodal abstractness of concepts to learn better image/text correspondences. They reported an improved retrieval performance through the introduction of concreteness and imageability in word embeddings of multimodal datasets. In a similar sense, Hewitt et al. [6] used the concreteness of concepts across multilingual image datasets to improve the results of translations.

2.3.3 Tasks quantifying human perception

Various tasks are proposed related to human perception of videos and images, as discussed in a related survey paper by Constantin et al. [102]. Next to *visual interestingness*, this survey paper cited a variety of interesting ideas towards the visual understanding of human perception, including proposed tasks like *coping factor*, *affective value*, or *perceived complexity*. Media Memorability, and especially Video Memorability, has become a task in the recent Multimedia Evaluation (MediaEval) benchmark [36].

2.4 Applications

There are a number of researches using the imageability or concreteness of words either directly or indirectly by relying on semantic embeddings including this data. In the following, I will outline a number of interesting existing applications already using such ideas, as well as a number of promising applications that would profit from the outcome of this research.

2.4.1 Use-cases of word ratings

Some multimodal applications already made use of word ratings like imageability or concreteness scores to infer multimodal understanding. Tanaka et al. [3] used content concreteness of documents to find comprehensible documents, finding a positive correlation between concreteness and content comprehensibility. Otto et al. [100][101] analyzed the semantic relationships between image and text, predicting the relative abstractness level of an image-text pair for use in image captions. Zhang et al. [31] analyzed the implicit relationship of image and text for posters and advertisements. They looked at examples, where the depicted meaning of the image contents and the text slogan is *parallel equivalent*, *parallel non-equivalent*, or *non-parallel*, meaning whether they try to convey the same, or opposite messages to the viewer. Therefore, rather than comparing whether they share the same contents, it tried to correlate the intrinsic meaning of both image and text. In the evaluation, a mixture of nine different features from image and text, including Psycholinguistic metrics like specificity and concreteness, are analyzed. The work makes some interesting conclusions on which kind of feature decodes what kind of hidden information when it comes to intrinsic semantic relationships. Vempala and Preoțiuc-Pietro [9] do a similar-minded approach to categorize image-text pairs from Twitter. While they do not explicitly use word ratings but LSTM based word embeddings for the textual information, this parallel research is similar to that of Zhang et al. [31], so the research might benefit

from looking at psycholinguistic features. Li and Nenkova [69] used *imageability*, *concreteness*, and *meaningfulness* to predict sentence specificity. The proposed method can be used to estimate text difficulty or create simplified versions of a text.

2.4.2 Explainable AI

There is also the recently established new field called Explainable AI (XAI) [68]. In XAI, the goal is gaining a better understanding of the operation of black-boxed AI models. Therefore, the internals of neural networks are analyzed to see how the output of a classifier can be explained. The nature of a black-boxed model makes it hard to verify results, but also to debug misclassifications. As many multimedia applications use neural networks for processing of language, be it personal assistants or translation tools, additional insight on human perception can help to explain misclassifications or unnatural results. There have been analyses related to AI for the fields of aviation and medicine, where a faulty classification could be potentially fatal, as discussed by Holzinger et al. [71][72]. As a measurement for human perception and underlying semantics, a way to estimate imageability for a large word corpus could help in gaining a better understanding of black-boxed models involving vision and language. This field looks at the problem that recent machine learning, especially neural networks, are often black boxes. There is very little 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. 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 [70], 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.

2.4.3 Sentiment

Another typical use-case for Psycholinguistic features is sentiment and emotion analysis. There is various research on sentiment and emotion in multimedia applications [76], spanning visualization, datasets [82], and recognition techniques [75]. Here, the goal is to find the sentiment triggered when reading a certain comment, looking at a certain image, reading a certain news, and so on. For sentiment evaluation, there are datasets such as LIWC [61] and Empath [63], 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 in regard to human perception. As such, it has become the topic of regular workshops affiliated with both Multimedia [67] and Natural Language Processing conferences [62].

Chapter 3

Relative visual variety differences for concepts in a narrow domain

Chapter 1 discussed the quantification of the mental image as a problem that can be divided into the two tasks of relative and absolute measurements. In this chapter, an approach to estimate the relative visual variety by means of a data-driven method is proposed. In the core assumption discussed in Section 1.3, the mental image of a concept was compared to what people imagine when thinking about a said concept. Following, the composition of a mental image for a concept is a mixture of things one visually associates with it; The image for `vehicle` might contain `cars`, `boats`, `planes`, and other things. Some concepts such as `cars` might, however, be more prevalent than others, like `tank`, despite both being considered as `vehicle`. In this research topic, this train of thought is exploited. As a data-driven method, one set of images per concept is accumulated (from now on called *imageset*). The set of all imagesets for a domain is called a *corpus*. The visual feature space of each imageset is compared to those of other imagesets to give a relative ranking across concepts in a narrow domain. The most abstract term serves as a reference point to predict a relative ranking across this domain of words.

Based on existing taxonomies for languages like WordNet [45], a set of sub-ordinate concepts for every composite concept is collected. Figure 3.1 illustrates such a tree

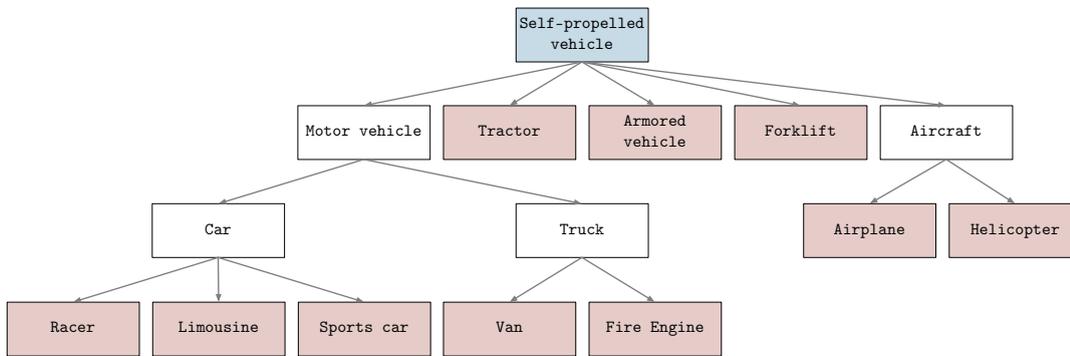


Figure 3.1: A simplified excerpt of the concept tree for self-propelled vehicle from WordNet [45].

for the concept `self-propelled vehicle` (the thesis refers to this example as `vehicle` from now on). An imageset for `vehicle` would contain cars, trucks, airplanes, and more. In the proposed method, the imageset for the concept in the root node in blue is a composition of images for the subordinate concepts in the leaf nodes in red. However, some subordinate concepts might be more prevalent than others, as many people might think of cars first when asked about vehicles. Following, in the proposed method, existing corpora are recomposed so that the ratio of images per concept should follow Web-based popularity, closely resembling the idea stated above. Clustering the visual feature space of these recomposed imagesets, relative variety differences between related concepts are calculated.

This chapter is structured as follows. Section 3.1 introduces the core idea and background of this research topic with Section 3.2 summarizing the contributions to the academic communities made through this research. Going into the details, Section 3.3 describes the proposed approach of visual variety measurements through cluster counting. For the approach to yield meaningful results, a well-balanced image corpus is necessary. Therefore, Section 3.4 proposes a method to construct such a corpus using Web-based popularity metrics as a weighting. Then Section 3.5 describes 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 3.6 shows the evaluation results, which are further discussed in Section 3.7. Finally, this research topic is summarized in Section 3.8

3.1 Motivation

In Section 1.3.1, research topic 1 was introduced as a method for estimating the related visual variety of concepts. In the following, these ideas are discussed in a greater detail.

In this research topic, 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. 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 (imageset) is created for each concept. As previously illustrated in Fig. 1.6, 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 imageset will look like. Imagesets of concepts in the higher-level of a concept tree 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 the score. Thus, how image corpora are composed is crucial for each measurement. For each concept, a well-balanced set of images resembling its common mental image is built, as shown in Fig. 3.2

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 seems 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 assumed to resemble it due to its crowd-sourced nature. Therefore, the core assumption of the proposed method is that the popularity of concepts on the Web approximates the general mental image of these concepts, and thus that there

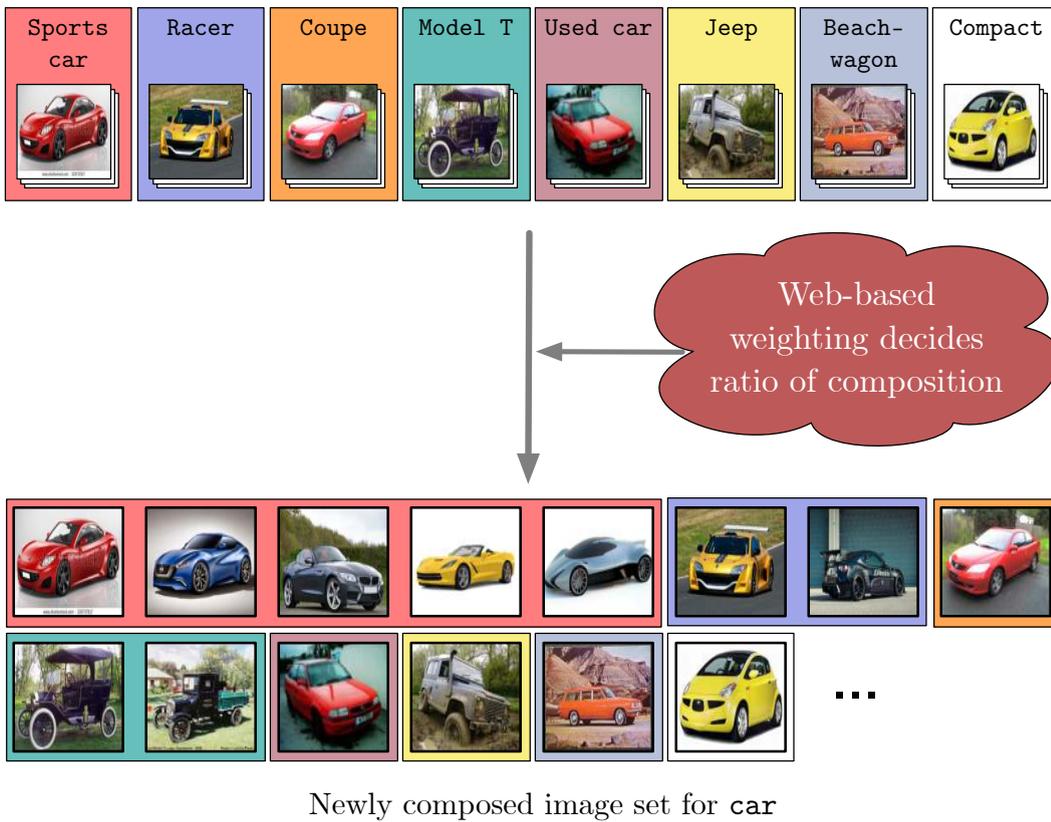


Figure 3.2: Creating a balanced imageset for car based on its subordinate concepts.

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 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.

3.2 Contributions

This section summarizes the novel contributions of the research topic described in this chapter. The idea of visual variety to quantize the mental image of a visual concept is a novel concept proposed as part of this research. For the estimation, the main focus of this Research Topic 1 is a data-based method which creates improved imagesets based on a recombination with Web-based weighting. To evaluate the experimental results, a ground-truth dataset of visual variety labels for a selection of 25 words is determined using a crowd-sourced survey.

3.2.1 Concept: Visual variety as a way to quantize the mental image of a visual concept

For this research topic, I propose the concept of *visual variety* as a means of measuring the semantic gap by approximating the common mental image of concepts. Comparing the size of feature space across different imagesets individually created for each concept, a measurement defining the perceived variety of such concept is calculated. There has been preliminary research on the idea of visualness [27] [41] comparing the entropy of Adjective-Noun Pairs (ANPs). Extending this idea, this research proposes a relative measurement of visual feature spaces between different concepts rather than different ANPs based on their visual characteristics.

3.2.2 Method: Image corpus recombination to adjust bias of existing image datasets

A major stepping stone when comparing imagesets of related concepts is the bias often seen in existing visual datasets. Following, comparing the visual characteristics can become very noisy, if the composition of the imageset does not reflect the expectation when mentally visualizing this concept in ones head.

To decrease the bias and thus to improve the composition of existing imagesets, a recomposition step is introduced in the proposed method. A Web-based popularity measurement is introduced as a weighting which decides the ratio of each sub-concept of a super-concept imageset. Following, an imageset for, e.g., `vehicles` would have a relatively high number of images of `cars` and a very low number of images of `tanks`, reflecting the expected composition of images when we think about the concept of `vehicles`.

Lastly, the number of clusters in the resulting visual feature space of each imageset is used as a relative visual variety score. For the experimental results, an unmodified baseline corpus is compared to recomposed corpora with different weightings, evaluating the viability of the proposed method.

3.2.3 Survey: Establishing ground-truth visual variety labels

As visual variety as a concept is proposed through this research, there are no existing ground-truth scores to compare the experimental results with. Therefore, there was the need for a ground-truth annotation to see how humans perceive each concept, in order to qualitatively evaluate the proposed method.

A collection of 25 concepts related to `vehicles`, spanning concepts like `car`, `aircraft`, and `boats` were manually selected. Using Social Media services Facebook ¹, Twitter ², and Reddit ³, a crowd-sourced survey has been performed, asking 150 participants on their perceived visual variety of these concepts. The survey was conducted using Thurstones' method of paired comparisons [47]. The responses were used to compute a ranking to which the proposed method can be compared with. More details on the survey are discussed in Section 3.5.

The computed ranking is used for the experimental results.

¹<https://www.facebook.com/>

²<https://www.twitter.com/>

³<https://www.reddit.com/>

3.3 Visual variety measurements

Distance measurements are commonly a direct comparison of two visual concepts [23]. 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 vision and language. Related work [27][39] analyzes 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. Inoue and Shinoda [22] analyzed, the visual relationship of terms within taxonomies. 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 vision and language. There is work regarding visual diversification [21], that aims for a large visual variety in image retrieval result sets. In the evaluation, their approach is 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 is 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 texts 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 [45] 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 research topic, 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. A set of images is defined as *balanced*, if there is a meaningful image composition which closely resembles a common variety of a concept. For every concept, an imageset 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 images in each concept's imageset. Thus, the ratio of how many images of each subordinate concept are within an imageset 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. 3.3, which shows the visual space of the concept **vehicle** as an example. For each concept,



Figure 3.3: Clustering the visual feature space of a concept (e.g., vehicle)

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

where $\#$ is the number of clusters and \mathbf{x} is a concept. For a concept \mathbf{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 imagesets if the same number of images is used.

3.4 Image corpus construction

Lexical relations within natural languages are commonly described using hierarchical structures. WordNet [45] provides a hierarchical collection of English words and terms. A collection of synonyms is called a *synset* and corresponds to a specific concept. For example, two separate synsets called *craft* could refer to the specific concepts of *handicraft* and *aircraft*. The hierarchy 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 [25] has a large corpus 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 are 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 data, this research 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.

3.4.1 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. As shown in Fig. 3.1, it is a *leaf node*, as there is no hyponym 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 imageset 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 imageset 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 biased set of images, as for ImageNet, the goal is to provide an overview of images necessary to grasp the coverage of its concept. In addition, 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 hyponyms of a concept can be a misleading 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.

3.4.2 Recomposition into a balanced corpus

The composition of images plays a crucial role for the perceived variety of the image corpus. For this, each non-leaf node imageset is recomposed 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

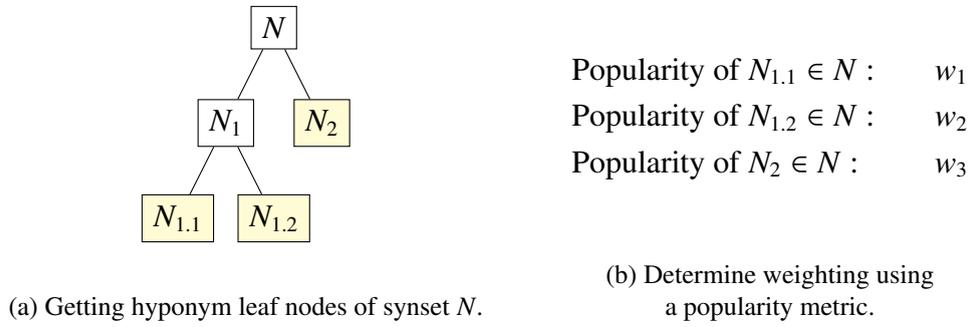
which are interchangeable but commonly have the same meaning when used in a related context (e.g. *cab* and *taxi*).

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 seems 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 [37], it is possible to crawl an approximation of the total number of 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 that a term or phrase appears.

Applying such a metric, a *popularity score* for each synonym for each synset is chosen. In Section 3.7.1, 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. Taking the average of multiple synonyms has the intrinsic advantage over taking a maximum popularity score that biased or noisy outliers are averaged out. The non-leaf imagesets are merged together using the previously determined ratio, as explained in Figure 3.4. This is believed to be superior to a crawling of non-leaf node images, as the composition of images would be uncertain and hard to validate.

3.4.3 Expanding the volume of the corpus

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



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

(c) Recomposing the image corpus. p is the function for retrieval of synset images.

Figure 3.4: Recomposition of the imageset for a synset.

obscure to have an influence on more abstract imagesets, they are removed from the hierarchy.

As the non-leaf node imagesets are composed from multiple leaf nodes, the amount of leaf node images becomes a major bottleneck. Extra images are crawled using Search Engine API [17][37] 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. In the experiments, the dataset for, e.g., *sports car* is increased by crawling for *sports car vehicle*. Note, that this modification is used for crawling additional images, but not for the popularity score described in Section 3.4.2. The full process of image corpus construction is visualized in Fig. 3.5.

Of course, Web-crawled approaches introduce a very high ratio of noise. Kennedy et al. [20] suggest a more than 50 percent chance of noise, even for dedicated image services like Flickr⁴. For Google Image Search [37], the ratio seems to be even

⁴<https://www.flickr.com/>

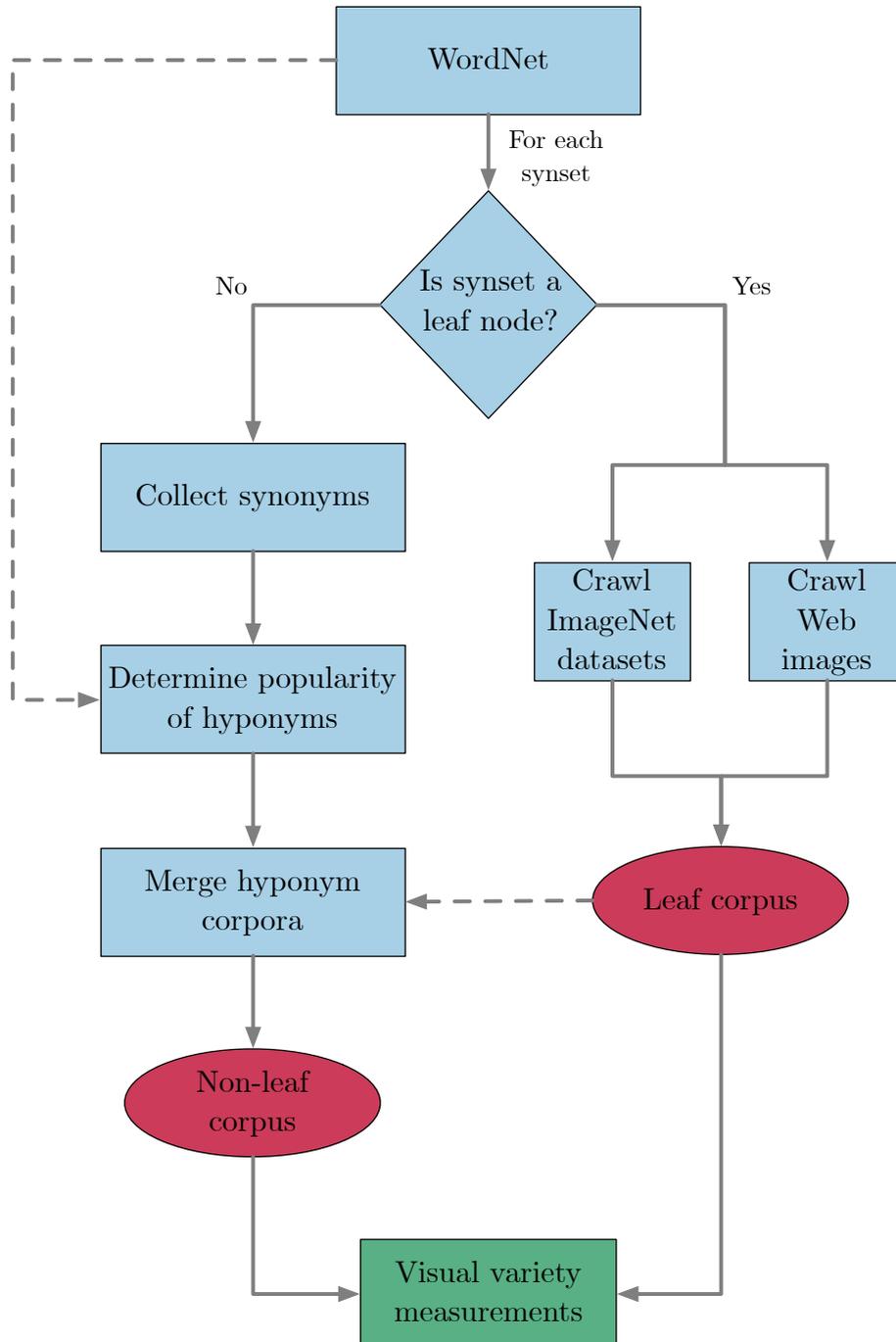


Figure 3.5: Flowchart of image corpus construction.

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 [29]. As there is a semantic relationship which corresponds to why the noise exists in the first place, removing noise images could also remove hidden semantics. Therefore, there is no further attempt to filter out noise images in this research.

3.5 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 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 my 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 is annotated with visual variety labels.

3.5.1 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 paired comparisons [47] 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. A survey was setup to conduct such an experiment for this research. On each page, a participant sees the name (e.g. "vehicle") and a short dictionary 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

Survey

Which concept **related to vehicles** has more visual variety?

jeep

A small, sturdy motor vehicle with four-wheel drive, especially one used by the military

jeep has more variety

sailing vessel

a vessel that is powered by the wind; often having several masts

sailing vessel has more variety

(a) Survey: Main part

How variant are these words?

Let's create a mental image for them, and think about it for a second...

animal



A lot of different looking animals exist.

cat



Similar animal in different situations...

(b) Survey: Tutorial

Figure 3.6: User interface for the crowd-sourced survey.

biasing the results in a predefined direction. To avoid confusion and misjudgement based on knowledge, all chosen synsets are commonly known terms.

For every paired comparison of concepts A and B, a participant can choose one of the following four options: “A has more variety”, “B has more variety”, “About equal”, or “I don’t know”. The user interface is shown in Fig. 3.6(a). 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 paired 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 3.6(b) shows an example of the tutorial page explaining the synsets `animal` and `cat`. Afterwards, the tutorial goes back to showing the participant the previous paired 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 participants would likely agree with 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.

3.5.2 Results

Over the course of two months, the survey has been promoted through Web and Social Media including Facebook ⁵, Twitter ⁶, and Reddit ⁷. Compared to solutions like Amazon Mechanical Turk (AMT) [56], this has the effect that mostly volunteers are participating in the survey. As participants are not paid, this can decrease the risk of spammers and thus improving the quality of results. Largely, a majority of answers seemed to take the survey diligently —most results match and people took a reasonable amount of time for answering each paired comparison. There were, however, a small number (around 5 percent) of dubious cases where people answered the survey suspiciously. Here, people evidently skipped most explanations and the time taken per paired comparison became significant outliers compared to others. As these answers 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 a crowd-sourced manner. While there was no restriction to native speakers, the participants were asked 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 node of the concept. The descriptions were sourced from Merriam-Webster’s Dictionary [49], Oxford Dictionary [50], 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 30 paired comparisons. Voluntarily, participants were able to extend the survey, in which case more unique paired comparisons would have been shown, but only one participant chose

⁵<https://www.facebook.com/>

⁶<https://www.twitter.com/>

⁷<https://www.reddit.com/>

to do so. Likewise, any participant was able to stop the survey at any point, in which case only the paired comparisons up to that point have been saved to the database.

In total, 158 people participated, answering 4,529 paired 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 paired comparison in average took 8.35 seconds. Out of all pairs, 87 percent 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 percent of problematic pairs without a majority, there were a couple of noticeable 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 ambiguous 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 confusion and nonhomogeneous results.



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



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



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

Figure 3.7: Examples for different corpus recompositions of the same synset.

3.6 Experiment

To evaluate the proposed method, corpora created with different popularity metrics as well as the unmodified baseline corpus are compared to ground-truth values obtained from the conducted crowd-sourced survey.

3.6.1 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 corpora 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 [44]. Leaf nodes with a too small amount of images (less than 100) 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 is 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 [37] and Bing [17] APIs are 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 is 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 3.7.1 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. 3.7. 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 Search-based (Proposed method 1) and Google Image Search-based (Proposed method 2) popularity metrics 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 corpus 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 corpus is an unmodified copy of the original ImageNet, there were no means taken to increase its amount of data.

3.6.2 Survey results

Based on the results from the survey discussed in Section 3.5, 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 [48]. For further steps, the ranking is 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 3.1 and Figure 3.8.

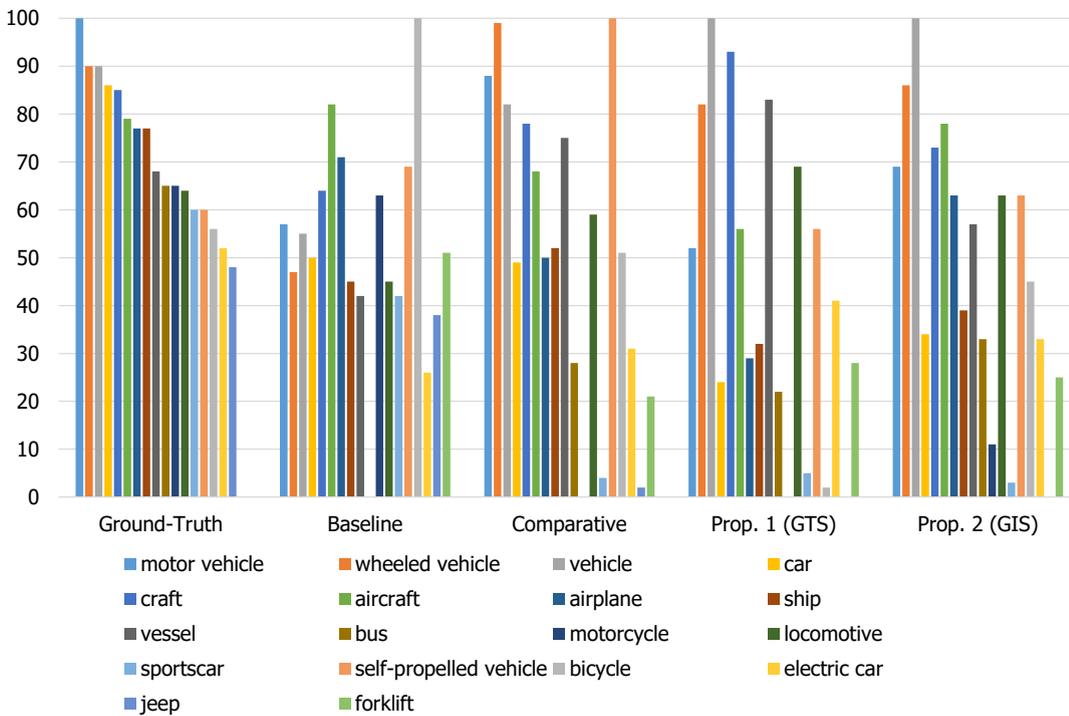


Figure 3.8: Visualizing the overall trend of each corpus.

3.6.3 Measurement results

The evaluation examines the data clustering of each imageset, as previously discussed in Section 3.3. For each synset, the number of clusters within the visual feature space of the synset’s images represents the visual variety.

The implementation uses OpenCV 3.2 [1] for feature extraction and distance measurements, and Scikit-learn 0.19.0 for clustering [14]. For each image, the visual features are extracted in form of a Bag of Words model using SURF descriptors [26][43]. A mean-shift clustering [16] 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 3.1 and Figure 3.8 show the ranking results for each corpus.

Table 3.1: Examples of estimated visual variety results.

Synset	GT	Baseline	Comparative	GTS	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

Table 3.2: Quantitative analysis of the proposed method.

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 Search weighting (Prop. 1)	0.56	14.89
Google Image Search weighting (Prop. 2)	0.73	9.01

3.6.4 Rank comparison

A comparison of the ranking generated for each corpus with the ground truth can be seen in Table 3.2. As metrics for evaluation, the Spearman Rank Correlation [57] and the Mean Squared Error (MSE) have been chosen.

The GIS-based proposed method 2 is leading the Rank Correlation with an improvement of 17.7 percent over the comparative method “Equal weighting” and 192 percent 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 3.1 and Figure 3.8, we 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 imageset 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 GIS-based 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 GTS-based proposed method 1 worsened the results, as it highly increased the MSE.

For both version of the Proposed method 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.

3.7 Discussion

The previous evaluations in Section 3.6 looked at how recomposed image corpora compare to a conventional corpus for visual variety measurements. It shows, that a recombination 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 recombination and obtaining a viable ground truth are discussed.

3.7.1 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 [37], where the maximum number 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 3.6.

Third, the Sketch Engine (SE) [38] 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 (Search Engine Optimization) keywords or Google PageRank, and solely relates on crawled text-only data. Lastly, the Corpus of Contemporary American English (COCA) [40] 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.

Table 3.3: Comparison of different Web popularity measurements.

(a) Distribution of the synset <code>truck</code>					(b) Distribution of the synset <code>car</code>				
Leaf node	GTS	GIS	SE	COCA	Leaf node	GTS	GIS	SE	COCA
<code>moving van</code>	22.8%	27.4%	2.4%	1.4%	<code>sports car</code>	32.5%	27.4%	45.7%	1.2%
<code>delivery tr</code>	9.6%	23.7%	1.8%	0.9%	<code>racer</code>	6.7%	9.2%	0.3%	2.3%
<code>pickup</code>	14.7%	10.9%	1.7%	44.0%	<code>model t</code>	24.0%	8.8%	0.8%	1.3%
<code>trailer tr</code>	7.1%	8.5%	2.5%	5.8%	<code>coupe</code>	2.3%	6.9%	3.5%	3.6%
<code>fire engine</code>	11.4%	6.8%	1.0%	2.6%	<code>used-car</code>	11.0%	6.7%	0.4%	1.8%
<code>tractor</code>	6.8%	6.0%	12.8%	26.8%	<code>jeep</code>	1.8%	5.0%	1.3%	6.4%
<code>police van</code>	9.8%	4.2%	58.4%	10.7%	<code>beach wagon</code>	2.2%	4.8%	2.5%	6.7%
<code>milk float</code>	1.8%	2.6%	0.3%	0.0%	<code>compact</code>	3.3%	4.5%	0.4%	11.0%
<code>transporter</code>	2.6%	2.1%	0.6%	1.6%	<code>cab</code>	1.9%	3.9%	3.4%	13.3%
<code>lorry</code>	1.9%	2.2%	7.8%	1.0%	<code>hatchback</code>	2.7%	1.2%	11.4%	1.1%
					<code>ambulance</code>	1.4%	0.6%	0.8%	15.9%
					<code>minivan</code>	1.3%	0.7%	8.5%	4.8%

In the following, the ratio found by each of these four corpora is compared. Table 3.3(a) shows the distributions for the synset `truck`, while Table 3.3(b) those for the synset `car`. For the synset `car`, the Web-based approaches often compose 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.

3.7.2 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`, which becomes 45.7 percent of `car` images for the Sketch Engine (SE) metric (Table 3.3b). 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 recombination, but this inevitably results in less variety for the recombination of parent nodes and thus some introduced bias.

3.7.3 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. 3.9. As seen, a tendency of the final results are determined rather quickly. Following, for future research it can be noted that it seems roughly 15 participants would be needed to get results closely resembling the final results. With more participants and the results getting more refined, the results for the proposed method gain a stronger lead.

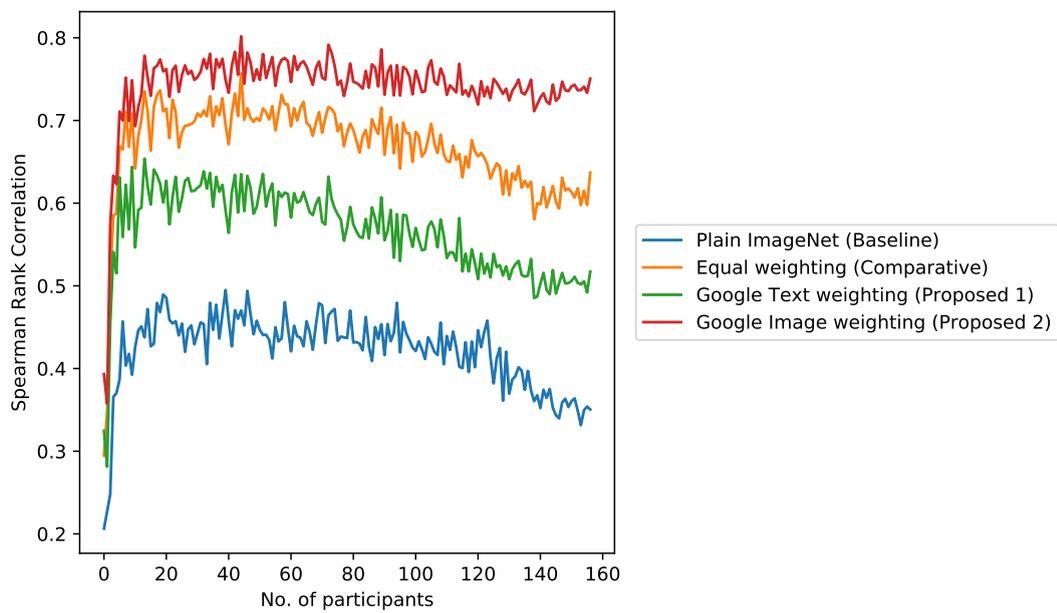


Figure 3.9: Stability of Spearman Rank correlation results for different ground-truths.

3.8 Summary

Chapter 1 introduced the concept of quantifying the mental image using image data. The problem was divided into the two tasks of estimating relative and absolute measurements. In this chapter, the first sub-task of estimating relative measurements has been tackled by means of a data-driven method proposed for relative visual variety measurements.

In this research topic, I proposed a method to measure the relative visual variety of terms using reconfigured imagesets modified to reflect Web-based popularity ratios. Web data is used to create and enhance an imageset for each term based on popularity in social media. The cluster counting method 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 image corpora with another, it shows that the correlation to ground truth highly depends on the used recombination. Compared to the baseline corpus, the recombination of proposed method 2 improved the measurements by 192 percent, showing very promising results in terms of understanding the relationship between vision and language.

The results showed good performance on a narrow domain for 25 words related to vehicles. Other domains, where the approach is considered to work, would be animals, plants, and so on.

Due to the data-driven nature of the proposed method, there are however some downsides which make it only feasible for a limited domain: The approach is tied to WordNet, meaning that a concept which has no meaningful hierarchy with hypernym/hyponym relationships is hard to recombine. This means, datasets for more abstract concepts such as *peaceful*⁸ are hard to create. Another limitation is the reference point used for the ranking: As the root term of the narrow domain is used as a reference point, the used algorithm of comparing clusters is somewhat tailored to narrow domains. A comparison of the number of clusters between vehicles

⁸While there is a WordNet synset for *peaceful*, adjectives do not possess hypernym/hyponym relationships.

and cars is reasonable, while a comparison between vehicles and pizzas has no meaning. As such, the proposed method can not deal with arbitrary concepts or a selection of words which goes beyond a specific domain.

Chapter 4

Absolute visual variety estimation for arbitrary concepts

Chapter 1 discussed the quantification of the mental image as a problem that can be divided into the two tasks of relative and absolute measurements. In this chapter, an approach to estimate the absolute visual variety by means of an algorithm-driven method is proposed. In the field of Psycholinguistics, there are existing dictionaries putting words on a Lickert scale regarding how they are perceived by humans, e.g., scoring them from one to seven. One of various such word ratings is called *Imageability*, describing whether a word is easy or hard to imagine. Visual variety is considered as a similar measurement, looking into how datasets for different terms differ in their feature variety.

As the second research topic discussed in this thesis, the idea of visual variety is employed for imageability estimation. While Research Topic 1 looked into relative differences between related terms, e.g. `car` and `vehicle`, imageability as a concept strives for absolute results —on a scale from vague words like `something` over abstract words like `peaceful` to concrete words like `car`.

This chapter is structured as follows. Section 4.1 introduces the motivation and background of this research topic with Section 4.2 summarizing the contributions to the

academic communities made through this research. The core assumption of how image data crawled from the Web correlates to the human perception of imageability is discussed in Section 4.3, together with the proposed method and the used mixture of low-level and high-level visual image features. For the evaluation, a large dataset composed of 1,000 words with imageability scores is prepared, discussed in Section 4.4. Section 4.5 analyzes the proposed method through four experiments, looking at the choice of image features, the choice of regressors, dataset size, and how the choice of visual features affects the performance for lowly or highly imageable words. Section 4.6 discusses the found results, as well as some implications for future applications. Finally, this research topic is summarized in Section 4.7

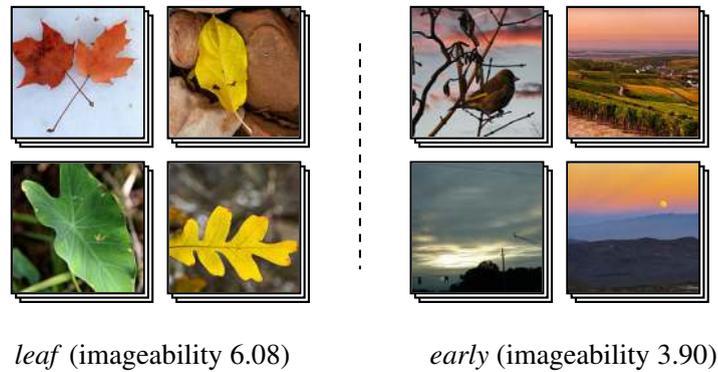


Figure 4.1: Core concept of word imageability.

4.1 Motivation

In Section 1.3.2, Research Topic 2 was introduced as a method for estimating the absolute visual variety of concepts. In the following, these ideas are discussed in a greater detail.

Imageability is a concept originating from Psycholinguistics [28]. It quantizes the human perception of words on a scale from, in layman’s terms, abstract to concrete. As a metric, it describes the ability to conceptualize a term as a mental image. A word with a high-imageability score is usually something rather concrete, for which the average person has an instant and rather clearly defined mental image, like *car* or *pizza*. In contrast, a word with a low-imageability score is something rather visually unclear, which is more of a concept than an actual object, like the word *transportation* or *nutrition*. As a consequence, imageability of words also correlates with *text difficulty*, as abstract, unclear words are often harder to grasp. Research in Psychology shows, that this relationship of language and imageability has further implications for language acquisition for children [51][55], language understanding [54], and the use of grammar [11]. The concept of imageability, along with example images for different imageable words, is visualized in Figure 4.1.

It seems natural to put this research in a Natural Language Processing (NLP) context, and use it for multimodal applications. While there have been multimedia applications which include Psycholinguistic concepts, there are various opportunities for

other fields to include such metrics, too. It is also recently used as a complementary feature for sentiment research [62][67], but found its way into recent multimodal research using text and image [31]. For automatically generated image captioning, such metrics could be used for quality assessment, both in terms of understandability, analyzing how text and image complement each other, or assessing the accessibility of texts.

Unfortunately, existing dictionaries used in Psycholinguistics are typically created through labor-intensive experiments. This can range from annotations by hand from test subjects in academic studies, to crowd-sourced surveys using online platforms like Amazon Mechanical Turk ¹. While there are a number of dictionaries for many languages, they tend to be rather small, especially compared to the word corpora of natural languages.

In this research topic, I propose a method using image-based data-mining to estimate the imageability of words. The core assumption is, similar to that of Research Topic 1, that imageability is a quantization of mental image of a certain word, describing how the society perceives it, and intrinsically reflected by images crawlable from the Web and Social Media.

Therefore, in this method, a large imageset is crawled for each word for which the ground-truth score for imageability is available. Next, a data-mining approach using a set of visual features is applied to all images. The visual features are selected to express a variety of visual characteristics spanning from very abstract to very concrete. Therefore, this approach includes a mixture of both a set of low-level, machine-based features, and a set of high-level features closer to the human description of images. For each word, a similarity matrix to describe the structural resemblance of all images in the same imagesets, is calculated per visual feature. Last, a model is trained to regress the imageability for unknown words. The model is evaluated using a series of testing data. In the experiments, first, the general performance of the proposed method in comparison to the previous work as well as a text-based method is

¹<https://www.mturk.com/>

evaluated. Then, the feature selection gets a closer inspection, to investigate which features can excel for which type of word. Finally, some implications following the results of each experiment are discussed.

4.2 Contributions

This section summarizes the novel contributions of the research topic described in this chapter. The core idea of this research topic is to employ the measurement of visual variety characteristics to the estimation of the word imageability as defined in the field of Psycholinguistics. For this, the approach uses a mixture of six visual features contributing to different parts of visual perception each complementing the overall image of a concept.

4.2.1 Concept: Visual variety for the estimation of imageability

Visual variety introduced in Chapter 3 proved to be a valid measurement for the perceived diversity of relative concepts in a limited domain, evaluated with a crowd-sourced ground truth. To test the core assumption regarding the mental image and Web-crawled images stated in Chapter 1, I wanted to test the idea as an absolute measure for arbitrary concepts in a larger-scale dataset involving a variety of domains.

In the field of Psycholinguistics, there are existing dictionaries putting words on a Lickert scale regarding how they are perceived by humans. For *imageability*, the rating describes whether a word is easy or hard to imagine. Visual variety is thought to be a similar measurement, looking into how datasets for different terms differ in their feature variety. As such, this chapter evaluates the usage of visual variety as a method for estimating the imageability of words. For the experiments, a dataset of 586 words, consisting both highly imageable and lowly imageable words is selected and evaluated to verify whether the method can be used for imageability estimation.

4.2.2 Concept: Mixture of low- and high-level visual features to complement semantic information

The main goal of the proposed method is to extend the algorithm used in Research Topic 1 with a more exhaustive set of visual features. With the imageability being roughly related to the abstractness of words, lowly imageable words will usually be rather abstract concepts with very vague visual characteristics. As an example, lowly imageable words like `peaceful` or `something` can usually not be depicted with a single object, but are rather a collection of ideas or concepts. In contrast, highly imageable words are often connected to existing objects like `car` or `leaf`, which have specific visual characteristics attached and can be more easily trained for a vision model to be detected.

Following this thought, this research topic uses a collection of six visual features, each contributing to different visual characteristics. The first set of three visual features represent patterns and colors, specifically not detecting objects but global attributes across the images. The second set of three visual features use object detection with a pre-trained neural network to represent higher level characteristics.

The experiments are outlined in a way to evaluate which set of features works better for which sub-group of words. This tests the hypothesis of whether, e.g., higher level features work better for concrete words, while lower level features work better for abstract words.

4.3 Imageability estimation

In this research topic, I propose a method to estimate the imageability of words using visual feature mining on Web-crawled images. The core assumption is that there is an intrinsic relationship between imageability scores and the perceived world around us. This relationship is considered to be reflected in image data on the Web, due to its crowd-sourced nature. While this can be both biased and subjective, photography and images on Social Media somewhat captures how we see the world around us. A large set of images related to a certain word will thus describe how the word can be visually represented in different ways, what situation it is commonly in, what common backgrounds (or varying backgrounds) for the said concept exist, and so on. This correlates to the mental image we have of the same word, and its imageability.

In Chapter 3, the relative *visual variety* of words in a narrow domain was considered. For this research topic, the focus is shifted from variety gaps within related words to general-purpose imageability estimation. The method of clustering local descriptors is prone to noise, as too many unrelated images will often connect clusters. When comparing *car* with *sportscar*, the clustering-based approach can spot the difference of variety, but comparing *car* to *pizza* will have trouble to find a reference point for comparison. In imageability estimation, both words would be similarly concrete. Thus, in this Chapter a more sophisticated method using a cross comparison of similarity between all images in the dataset of a word is proposed. Additionally, to successfully capture the characteristics of various concepts, four additional visual features are introduced. Lastly, a model is trained to predict an imageability score from the cross-similarities using ground-truth annotations from Psycholinguistic dictionaries consisting of common words from various domains.

4.3.1 Approach

Let's assume an existing dataset with imageability scores attached. For each word, a sufficiently large number of images from crowd-sourced origin is needed for the

data-mining to work as expected. *Imageability* is described as a numerical rating on a scale between rather concrete (usually high scores,) and rather abstract words or concepts (usually low scores.)

Concrete words are easy-to-grasp concepts, which are very imageable, but lack a variety. Think of the word *car*; while there is a large variety of different cars, most of them look fairly similar in their fundamental shape, form, and choice of colors. Furthermore, the *situation* a car is in would usually be very similar —A street, or scenery, but very rarely in the middle of the rain forest, or in the air (like a plane would be, on the other hand.)

Abstract words, in contrast, are often much less imageable. They tend to have a much higher visual variety, just through the nature of them being usually not objects, but atmospheres, situations, or concepts, on their own. Therefore, they cannot usually be described with single images, and images of the same abstract word will look very different from another. For example, the dataset for the word *approach* would probably contain many technical figures, but its visual characteristics are not well defined.

The proposed method exploits these visual characteristics. Words with high-imageability scores are expected to have high similarities across all their images. In contrast, words with low-imageability scores are expected to have significantly lower similarities across their images.

Using a variety of visual features (Discussed in Section 4.3.2,) a similarity matrix is built. For each visual feature, one histogram describing it is computed for each image. By cross comparing all images, the similarity of all histograms is calculated and inserted in a matrix of size $n \times n$ for n images. For a high number of images, the similarity matrix reaches a high dimensionality, which makes it hard to train a model with the similarity matrix as input. Furthermore, the similarity matrix changes with the order of processed images, despite the order having no meaning in itself. To solve these issues, the eigenvalues of the said similarity matrix are computed. The

eigenvalues contain the characteristics of the similarity matrix, meaning that the visual characteristics of low-imageability words' visual features vs. high-imageability words' visual features are also encoded in them. Meanwhile, a sorted set of eigenvalues has a significantly smaller dimensionality than the matrix, and it is invariant to changes in the order of images.

Lastly, a model is trained to regress imageability, using the previously calculated sets of eigenvalues as input. Existing imageability annotations from Psycholinguistic dictionaries serve as ground-truth scores.

The step-by-step algorithm is shown in Figure 4.2 and further described in Algorithm 1.

4.3.2 Feature selection

To sufficiently encode the visual characteristics of each imageset, the analyses need to look at visual features from multiple angles. Computer vision and object detection algorithms traditionally focus on low-level representations of visual characters. Patterns, edges, and color spaces are encoded and represented in forms of feature vectors. While this is important for many parts of computer vision, it also leaves human perception of concepts out of the image. For a human, the situation or actual contents of a picture is often more important than a global gradient description. Low-level features also do not contain actual meaning, if not trained against ground-truth data.

Therefore, the proposed method looks at the problem from two angles. First, low-level features are analyzed to have a general description of the scene and objects. This will furthermore relate to how humans perceive colors and contrasts, which are important parts of the core assumption of imageability. Second, high-level features are extracted using pre-trained models from Computer Vision and Multimedia applications. Here, I am interested in the actual image contents and compositions. The features are used to complement the visual feature representations in *what* and *how*

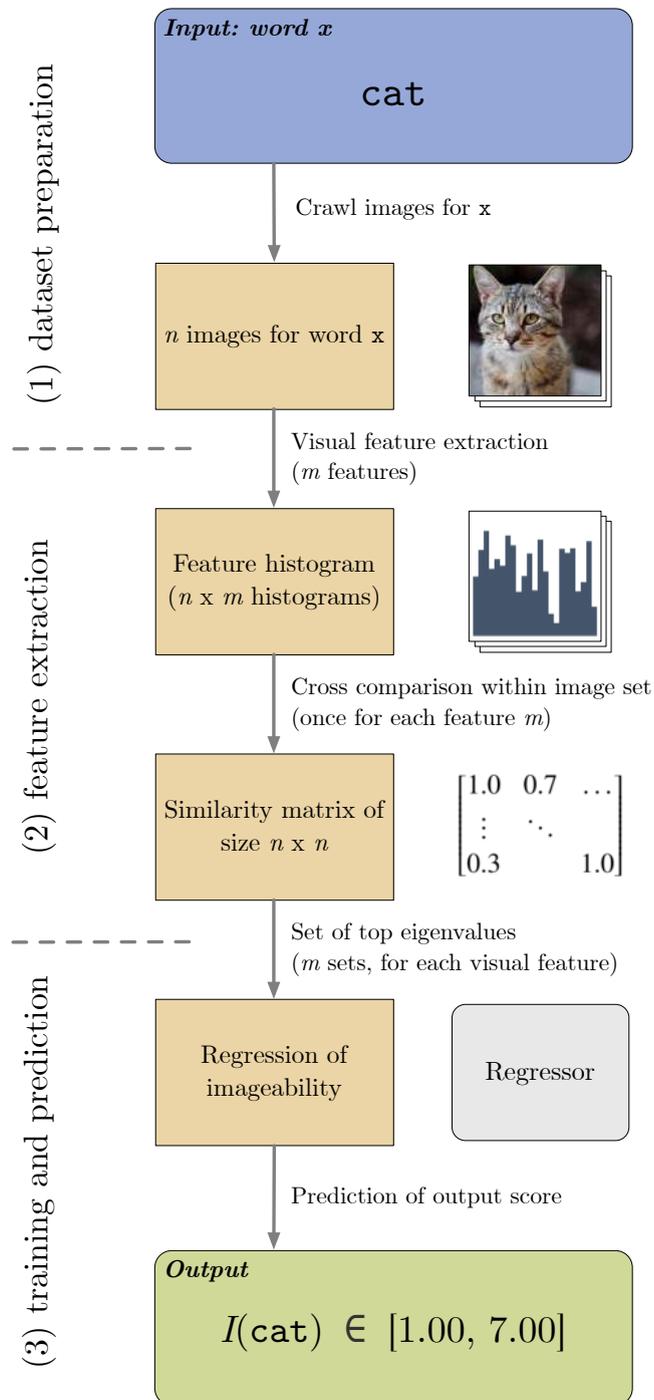


Figure 4.2: Flowchart of the imageability estimation process.

Algorithm 1: Pseudo-code for the proposed method.

input : Word

output: Imageability label

```

1 (1) data preparation;
2  $images \leftarrow$  image dataset;
3  $words \leftarrow$  psycholinguistics dataset;
4  $features \leftarrow$  list of visual features;
5 for  $image \in images$  do
6    $image\_text \leftarrow$  read textual metadata of  $image$ ;
7   if  $image\_text \cap words \neq \emptyset$  then
8     for  $word \in image\_text \cap words$  do
9        $images_{word} \leftarrow image$ ;
10    end
11  end
12 end
13 (2) feature extraction;
14 for  $word \in words$  do
15   for  $image \in images_{word}$  do
16     for  $feature \in features$  do
17        $images_{word,feature,image} \leftarrow$  extract visual features;
18     end
19   end
20   for  $feature \in features$  do
21      $similarity\_matrix_{word,feature} \leftarrow$ 
22     cross comparison similarity for all in  $images_{input,feature}$ ;
23   end
24 (3a) training;
25 for  $word \in words$  do
26    $X_{word} \leftarrow \prod_{i=1}^n$  Eigenvalues of  $similarity\_matrix_{word,i}$ (for  $n$  features);
27    $Y_{word} \leftarrow words_{word}$ ;
28 end
29 train regression model  $Y$  on  $X$ ;
30 (3b) prediction;
31  $X \leftarrow \prod_{i=1}^n$  Eigenvalues of  $similarity\_matrix_{input,i}$ (for  $n$  features);
32 predict  $Y$  from  $X$ ;
33  $output \leftarrow Y$ ;

```

things are displayed in each image, while putting the actual technical details (e.g., low-level details) to the side.

4.3.2.1 Low-level features

Low-level features represent the visual characteristics of each image *how a machine would describe them*. They encode local and global trends of edges, colors, and gradients of the processed image. While these are important characteristics and the basis for object detection and scene understanding, the actually encoded patterns do not possess much of a meaning on their own. In the experiments, the following low-level features are used:

Color distributions. The color distributions are captured as one visual feature. In context of imageability, this feature can encode the mood and the atmosphere of each image through the overall distribution of used colors. The atmosphere of a concept could be captured by finding reoccurring color patterns like *warm* or *cold* colors. Furthermore, this feature encodes information related to visual adjectives like *yellow* or *bright*.

Global gradient descriptions. Global features are important for scene analysis. They are, among other use-cases, prominently used for Web-retrieval engines. Based on an encoding of gradients, and their orientation, the feature representations give information on global pattern distributions of the images, such as how noisy an image is to the eye, whether there are many objects, and contrasts.

Local gradient descriptions. Local features are often used for object detection, as they can be used to distinguish the visual characteristics of different objects. In a sense, they decode the patterns of an object, and what makes it look like the object. In combination with a Bag-of-Visual-Words (BoVW) [26] model of the local gradient descriptors, it creates a histogram encoding reoccurring visual patterns within the

image. While this sounds more high-level than just edges, it is a different level of abstraction than actual high-level features, as the found patterns do not necessarily share meaning.

Actual implementation details of the feature extraction can be found in Section 4.5.1.

4.3.2.2 High-level features

High-level features look at the visual characteristics of each image *how a human would describe them*. While colors, contrasts, and edges are also part of how humans see objects, they have few actual meanings in themselves. The actual meaning comes from associating pattern recognition with ground-truth labels, which a model can be trained to find, but is not an actual part of the visual feature representation. In the experiments, the following three characteristics of high-level representations are investigated:

Image theme. First, the image theme is the overall setting of an image. Examples of this could be: *indoor*, *landscape*, or *architecture*. This is not an actual description of displayed objects, but rather the situation or scenery where all the displayed objects are in. The setting of an image plays a large role for similarity of images, as it is largely an encoding of backgrounds, which are often the largest part of each image in terms of surface area.

Image contents. Second, the image contents are actually displayed objects in the scene. A scene of two dogs and their owner in front of a crowded street might contain the objects: dog 2, human 1, cars 3, and so on. An object frequency along with an object description gives additional insights of the nature of each image. Because two images, one of a black small nude cat and one of a white fluffy cat, are perceived rather similar to humans despite having different colors or patterns, a high-level representation of image contents is needed.

Image composition. Third, image compositions give another insight on how important things are for the scene. Images with a certain object in the center of an image might directly relate to this object, while the same object in a corner of another image might just be part of the scenery. Furthermore, concrete, high-imageability words, might correlate to objects being in the center, while abstract, low-imageability words, might show other characteristics or general trends.

Pre-trained models are used to encode these characteristics and to describe them in the form of likelihood histograms. The resulting histograms are then used in the cross-comparison step proposed in Section 4.3.1 above. Actual implementation details in which models are used for the evaluations are given in Section 4.5.1.

4.4 Dataset construction

In this research, two types of datasets are employed. The first is a dictionary with English (language) words and 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.

4.4.1 Imageability dictionary

There are a number of imageability or concreteness dictionaries in different languages, including English [30][58], Indonesian [59], and Cantonese [60]. As described before, imageability dictionaries try to quantify the human perception of words. The most common scale is a seven-level Likert scale, averaging the perception over all test subjects. Level 1–3 words would be things where one can not grasp a mental image to describe it. In layman terms, when talking about nouns, it might be a rather abstract concept, like *peace* or the word *abstract* itself. It could also be a conjunction, which are naturally hard to visually image, like *because*. A level 5–7 word on the other hand is something rather concrete, which is easy to grasp. It could be a *dog* or the color *red*.

Datasets for imageability are commonly created by hand. Using crowd-sourcing or surveys, a pre-selected set of words is judged by each test subject. It could be measured using paired comparisons, which might arguably lead to more accurate results. However, the sheer amount of labor involved in this process results in most studies using Likert scales instead.

For evaluating the proposed method, in this thesis, the English language is used. Concretely, the datasets by Reilly et al. [58] and Cortese et al. [30] in combination are used as a baseline. These datasets provide the results as a Likert scale score averaged over all test subjects, in the range of [1.00, 7.00]. There is no significant overlap nor contradictions in both word corpora. Furthermore, while the former is

only composed of nouns, the latter includes other parts-of-speech. In case of overlap, the average of both dictionaries is taken.

While there are other datasets, combining a large number of different datasets might result in incomparable results, as it is unclear whether all experiments have been conducted in the same way. The popular, but also rather dated, MRC database [64] has not been used directly, despite it being larger than the previously cited sources. However, the first dataset used [58] is a modified version of the MRC database. It focuses on the high- and low-end of the spectrum, removing large parts of mid-Imageability terms from the original MRC database. In that process, they also filtered out obscure and uncommon terms, making for a cleaned-up fork of the MRC database.

Lastly, while Likert scales are very common in Psychology, Computer Science is used to either percentual results, or a normalized scale of $[0, 1]$. Therefore, for pure understandability of the evaluation results, the interval of $[1.00, 7.00]$ is normalized to $[0, 100]$ in this thesis.

4.4.2 Imagesets

For the image data, Social Media platforms are crawled for each word. The whole process of dataset acquisition is shown in Figure 4.3. The noisy Web-based origin ensures a composition which comes close to how a human perceives the concept. The simplicity of direct crawling, on the other hand, ensures that a larger number of images for a much larger number of words can be retrieved. Therefore, the proposed algorithm from Section 4.3 can be evaluated with a large number of words, also testing the stability of its predictions for different scales of datasets. As the proposed dataset creation method does not rely on WordNet, it implicitly groups ambiguous terms, and can be used for terms not available in the WordNet hierarchy, or is insufficient (e.g., there are multiple levels of hierarchy with no siblings.) Lastly, it comes without extra post-processing or manual labor needed for recomposing the dataset.

Using the imageability data described in Section 4.4.1 as a basis, a large number of images for each word with imageability annotation is crawled. As a source for the images, the YFCC100M [32] dataset is used, which is crowd-sourced based on the US photography social media platform Flickr ². It consists of 100 million images posted to Flickr up to 2014, annotated with various text-based annotations like a title, a description, user taggings, and more. The dataset also comes with 1,570-class visual concept annotations. This can be used as a high-level feature on its own and will be discussed later. Here, the images themselves are used for visual feature data mining. Furthermore, the text-based annotations are used to identify the relationship between images and words.

For each image, if a word from the imageability dictionary is contained in one of the text-based annotations (title, description, or user-tagging,) the image and the word are considered as related to each other. Thus, the YFCC100M dataset is crawled, looking for images where entries from the imageability dictionaries appear in the text annotations. In case of multiple related words, the image is flagged to be part of the imageset for each word.

To not bias the proposed method with different similarity matrix sizes, an equal number of images is used for every word. As the frequency of images for different words varies, many words are harder to crawl than others. For each word, the first n images retrieved in the crawling process are used for the evaluation. Furthermore, there is a large amount of noise and mis-classifications, which is natural for crowd-sourced Web-based data. Noise, like unrelated images, is expected to be averaged out if the number of images is large enough. For abstract words, the noise ratio is naturally much higher, as it is hard to put a concrete label on very abstract words. This characteristic helps the proposed method, as a high noise ratio results in a low cross-similarity between images and thus naturally produces the expected similarity matrix for abstract terms. The noise in lowly imageable word datasets is also shown in Fig. 4.5 in the next section.

²<https://www.flickr.com/>

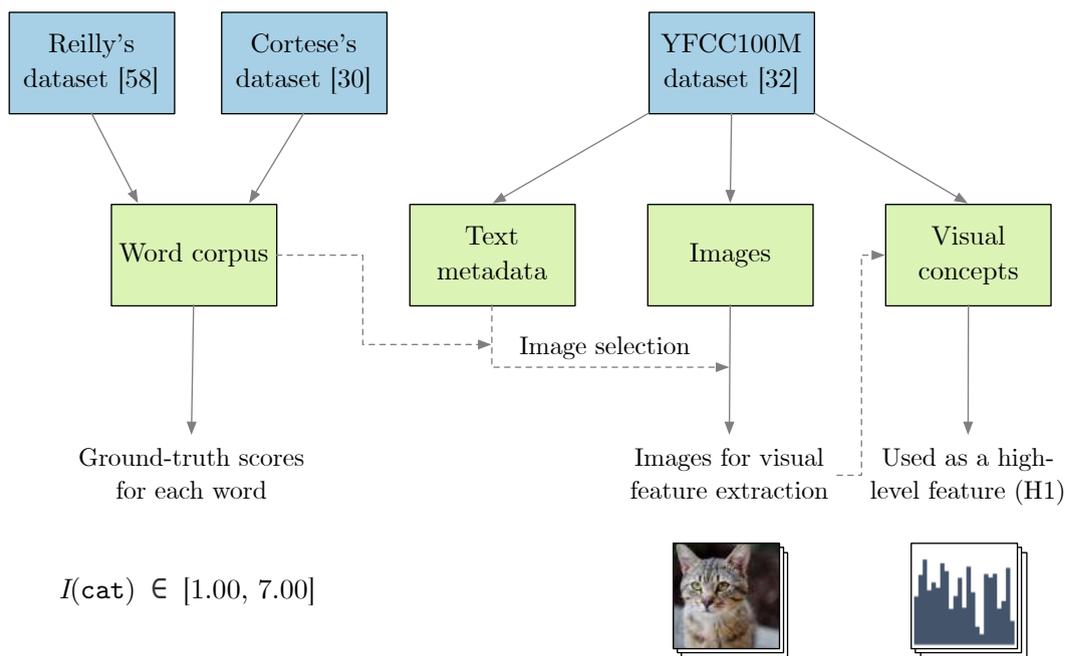


Figure 4.3: Flowchart of dataset acquisition of visual data for words and their imageability labels.

4.5 Experiment

The goal of this research is to estimate imageability scores using data mining on visual features of crowd-sourced images. The analyses use a variety of low-level and high-level features to provide a view on the visual characteristics from various angles.

In the following, five experiments conducted using a large Web-crawled dataset for imageability estimation are outlined. After discussing details on the environment of the analyses, first, the results when using different visual features are presented. Then, the dataset size, and how a larger number of images can influence the resulting error are analyzed, as well as how the choice of the regression model makes a difference to the proposed method. Lastly, two experiments will analyze which feature excels for which kind of words, both considering low-imageability vs. high-imageability as well as different parts-of-speech.

4.5.1 Feature selection

The evaluations use a combination of seven different visual feature sets. First, three visual features will encode the low-level visual information of each image:

(L1) The *HSV color feature* encodes the color distribution in the HSV color space. For the color features, it results in the best prediction performance for experiments when using 36 bins for the Hue and Saturation axes each, resulting in a 72-dimensional histogram for each image.

(L2) The *SURF feature* uses the SURF local feature transformation [43] to generate a Bag-of-Words model [26] using k -means clustering. SURF is a common feature used in object detection or reconstruction. The resulting 4,096-dimensional histogram describes the occurrence of visually similar sub-regions based on gradients.

(L3) The *GIST feature* uses the GIST descriptor [33] commonly used for scene analysis. Based on this global gradient encoding, a 960-dimensional histogram is generated for each image.

Next, four high-level features complement the low-level features above to provide additional information closer to human perception:

(H1) The *Image theme* feature captures the general concept of each image. In the following experiments, the YFCC100M-based autotaggings provided in the dataset (as shown in Figure 4.3) is used. The taggings include concepts like *inside*, *nature*, *architecture*, and more. The resulting histogram is composed of 1,570 classes, based on the probability of each concept being related to the image.

(H2) The *Image content* feature encodes objects in each image. In the following experiments, the pre-trained model YOLO9000 [65] is used to detect concrete objects found in each image. The frequency histogram is based on the number of detected instances for each class. The model YOLO9000 was specifically chosen because of the large number of classes, as newer versions of YOLO come with a substantially smaller number of classes. The 9,418-classes proposed in YOLO9000, however, turned out to be too many for a proper histogram comparison. To establish a middle ground, WordNet [45] is used to group classes along their hypernyms. Each class of YOLO9000 corresponds to leaf nodes in the WordNet hierarchy. They are combined in a bottom-up fashion, resulting in a dimensionality reduced to 1,401-classes after merging three levels of hypernyms.

(H3) The *Image composition* feature encodes the location of objects in the image. Again, YOLO9000 is used to detect objects within each image. Using an overlapped $n \times n$ grid, a histogram describing the number of objects within each grid cell is generated. In the following experiments, the actual value of n is set to 10, resulting in a 100-dimensional histogram.

Each feature is used to calculate a similarity matrix as outline in Section 4.3.1. The eigenvalues of the similarity matrix are used as input for the regression. If sorted by size, the top eigenvalues contain the majority of structural information of the matrix, and are least affected by noisy data. Thus, in the following experiments, the top 30 eigenvalues of each visual feature are used to simplify the training. This heavily decreases dimensionality and thus complexity for the training process, especially when working with combined features. For combined features, the resulting eigenvalues for each feature have been concatenated before inserting them into the regressor.

For all implementations, Python 3.7 and OpenCV 3.20 [1] is used. For YOLO9000, the Python implementation YOLO3-4-Py [13] is used. For histogram comparisons, the default normalized cross-correlation metric from OpenCV [1] is used:

$$d(H_1, H_2) = \frac{\sum_I (H_1(I) - \bar{H}_1)(H_2(I) - \bar{H}_2)}{\sqrt{\sum_I (H_1(I) - \bar{H}_1)^2 \sum_I (H_2(I) - \bar{H}_2)^2}}, \bar{H}_k = \frac{1}{N} \sum_J H_k(J)$$

where N is the total number of histogram bins.

4.5.2 Dataset

Following the process discussed in Section 4.4, datasets with ground-truth imageability annotations for up to 1,148 words (for 2,500 images each) and up to 587 words (for 5,000 images each) have been accumulated by crawling the first approximately one sixth of the YFCC100M dataset. The data can be increased for a bigger dataset and more accurate results, but I decided to stop further crawling at that point due to feasibility in processing time. As many words are much harder to obtain than others, the number of words available shrinks with the number of images wanted for the evaluation.

For the majority of evaluations, if not indicated otherwise, a dataset having 587 words with 5,000 images each has been analyzed. I found that this gives us a good balance of a sufficient number of images for data-mining, while having a sufficient

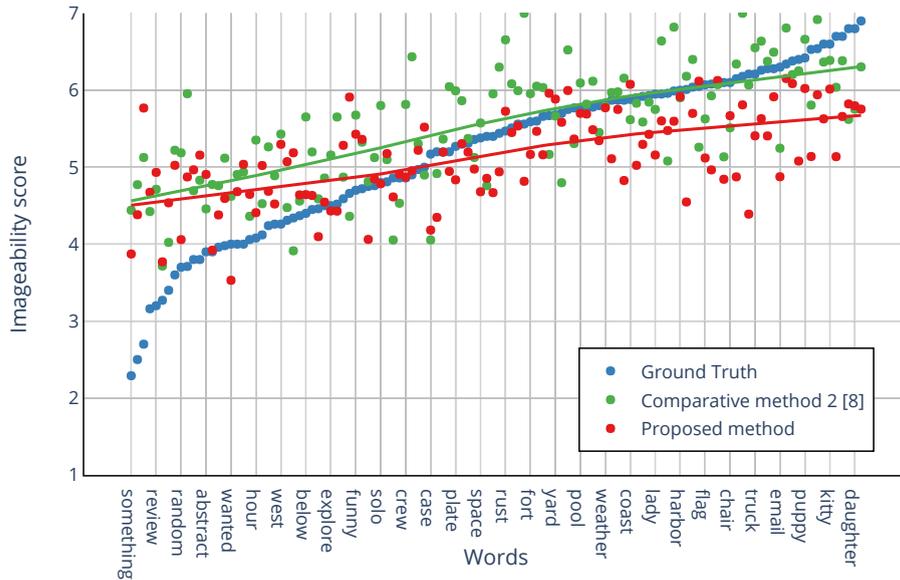


Figure 4.4: Scatter plot of predicted imageability scores.

number of training samples, and still being feasible in terms of processing power. It spans 501 nouns, 33 adjectives, 18 adverbs, 11 verbs, and 24 other parts-of-speech³. The average imageability in the training dataset is 67 (testing: 70) with a standard deviation of 20 (testing: 17). Thus, the dataset is biased towards highly imageable terms, but still contains lowly imageable terms. A scatter plot of the test dataset is also shown in Fig. 4.4, together with results for the proposed and comparative methods.

To investigate the effect of dataset size, the robustness against different numbers of words (thus, training samples,) and the number of images per word are also tested.

Example images from the created image dataset are shown in Fig. 4.5.

³Parts-of-speech are obtained using NLTK [44] and may thus have slight error due to ambiguities.

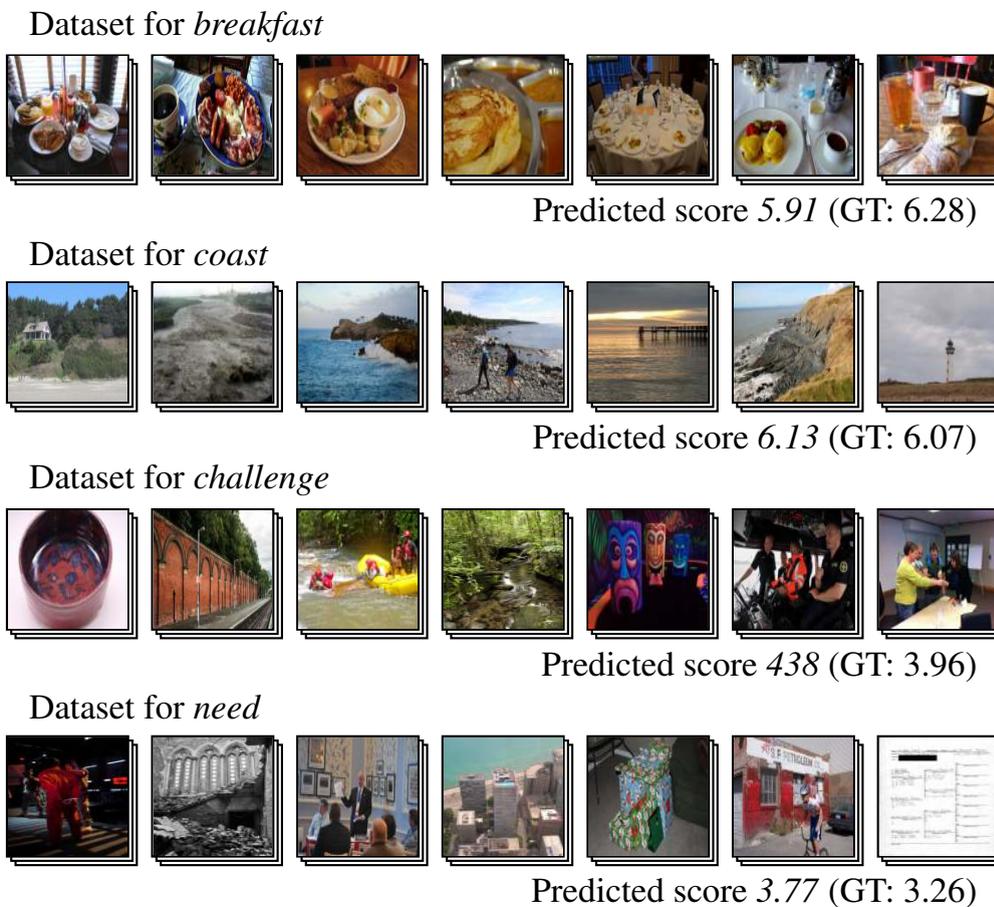


Figure 4.5: Example of image datasets and their predicted imageability scores.

4.5.3 Regression model

For training and evaluation, the datasets are split in 80 percent of words for training and 20 percent of words for testing.

The evaluations, if not indicated otherwise, use Random Forest [66] as the regressor. For comparison, an SVM-based regression and a shallow Neural Network have also been tested. The former two use Scikit-learn 0.19.0 [14], while the latter is implemented in Keras 2.0.6 [24].

The Random Forest uses 100 estimators. The SVM regression uses an RBF kernel with $C = 100$ and $\gamma = 0.001$. The Neural Network uses a shallow architecture with three Dense layers of 512 dimensions.

4.5.4 Evaluation metrics

All experiments are evaluated using two metrics: The first is the Mean Absolute Error (*MAE*) with the best result being 0 meaning no error compared to the ground-truth scores. The second is the Pearson's correlation coefficient (*Correlation*) with the best result being 1 (or -1) meaning a perfect ordering (or perfect opposite ordering) of the predicted scores.

In layman's terms, a low error but low correlation would mean that most predicted scores are rather close to their ground-truth scores, even if they would result in the wrong ranking order due to slight differences. As the ground-truth seven-level Likert scale score is chosen rather vague, and the dataset is furthermore biased towards highly imageable words, this results in many samples in the upper third of the results. Following this, it is possible to have a very low error but mixed correlation results.

The opposite would be true if there is an in-general good correlation between the predicted samples, but a couple of very strong outliers heavily influencing the MAE. This is true for some of the cases in the analysis of parts-of-speech, where the test dataset has a very small number of samples. Here, many results share the correct relative order of high- vs. low-imageability predictions among the same part-of-speech, but the error can be rather high as the training data consists of nouns, maybe unfit for the evaluated part-of-speech.

4.5.5 Results

In the first experiment, the proposed method has been evaluated on the dataset of 587 words with 5,000 images each. Table 4.1 shows the results for each feature selection. It reaches the best results with an error of 10.14 and a correlation of 0.63. The proposed method uses a combined vector with all high-level and low-level features except L3 (GIST.) When including L3, it results in a slight decrease to an error of 10.33 with a correlation of 0.62. Interestingly, H2 and H3 (both using YOLO9000 as there baseline,) have rather unfortunate results on their own, but can increase the

Table 4.1: Qualitative analysis for different sets of visual features.

Feature	Correlation (1 : best)	MAE (0 : best)
L1: Color histograms	0.53	11.30
L2: SURF + BoVW	0.54	11.48
L3: GIST	0.42	12.05
H1: Image theme (YFCC100M-based)	0.62	10.19
H2: Image content (YOLO9000-based)	0.43	12.55
H3: Image composition (YOLO9000-based)	0.25	13.98
Combined Low-level L*	0.60	11.03
Combined High-level H*	0.61	10.18
Combined (Proposed method; All)	0.63	10.14
Comparative method 1 (Visual variety, Research Topic 1)	-0.01	67.31
Comparative method 2 (Text data mining [8])	0.70	10.39

performance if combined with other features. This suggests that the visual features can complement each other well enough, as they each encode a different kind of visual characteristics. Overall, the combined high-level features perform better than the combined low-level features. While the H1 feature set, which is part of the YFCC100M dataset, performs good on its own, the combined proposed method with H1 excluded can still reach an error of 10.25 with a correlation of 0.63. This means that the method works similarly well for other datasets, where H1 features are not directly available.

For comparison, first, the cluster counting approach used in Research Topic 1 has been tested on the new dataset. In Research Topic 1, the variety of visual characteristics in a BoVW model is used to estimate a variety score for a dataset. It is closely related to the main assumption of this topic, although Research Topic 1 was not developed, nor evaluated, for the purpose of absolute imageability score estimation. The results show, that the performance of the proposed method for Research Topic 2 is superior for imageability estimation.

As a second comparative method, the predicted imageability scores has been compared to the method proposed by Ljubesic et al. [8]. Their method predicts imageability entirely based on text data-mining, while the proposed method exclusively

uses visual characteristics of images. This makes for an interesting comparison between different modalities. The results show a slightly better correlation of 0.70 for the text data mining method, but for the MAE, the proposed method wins with 10.14 versus an error of 10.39. These mixed results suggest that it would be beneficial for future work to combine both models and regress the scores using both textual and visual characteristics of multimodal datasets. However, due to the closely cluttered results of most of the testing dataset, a high correlation is very hard to achieve. The imageability for words closely neighboring on the Lickert scale is often very vague due to the seven-level nature of the ground-truth annotations. As such, the relative order might be very hard to decide, even for most human annotators. Following, I believe that the MAE is a better metric for this, more closely capturing the trend of predictions (highly imageable vs. lowly imageable) rather than the exact order of each result.

In the second experiment, to assess the stability of the results, the proposed method has been tested with different dataset sizes. In Table 4.2, the results for a varying number of words and a varying number of images per word are shown. For the varying number of words, the previously discussed dataset having 587 words and 5,000 images each has been used. The dataset has been split in 469 words for training and 118 for testing. For the reduced number of words, the model is trained with 321 (66 percent of training samples) and 156 (33 percent of training samples,) respectively. The results confirm that the error is sufficiently stable for different dataset sizes. They also show that the error decreases with the number of images. The complexity of the method scales linearly with the number of images for visual feature extraction, and quadratically for calculating the similarity matrices. The training time is negligible for the most part, but the pre-processing of visual features and the matrices is a major bottleneck. Using an RTX 2080 Ti (GPU-based visual features,) and a Xeon E5-2697 (CPU-based visual features and similarity matrices,) pre-processing the dataset for the 5,000 image/word dataset took several weeks. As the number of available words (i.e., training samples) for more images per word also further decreases, I did not look into larger dataset experiments.

Table 4.2: Qualitative analysis for different training dataset sizes.

Dataset		Correlation (1 : best)	MAE (0 : best)
Fixed #images	156 words / 5,000 images each	0.51	12.17
	321 words / 5,000 images each	0.62	10.58
	469 words / 5,000 images each	0.63	10.14
Fixed #words	469 words / 1,000 images each	0.54	11.27
	469 words / 2,500 images each	0.57	10.87
	469 words / 5,000 images each	0.63	10.14

Table 4.3: Quantitative results for different regressors.

Regressor	Top 30 eigenvalues		All eigenvalues	
	Correlation (1 : best)	MAE (0 : best)	Correlation (1 : best)	MAE (0 : best)
Support Vector Machine	0.13	14.82	0.11	14.83
Neural Network	0.61	10.82	0.60	11.11
Random Forest	0.63	10.14	0.63	10.17

In the third experiment, the regressor has been exchanged. Random Forest, SVM, and a shallow Neural Network were tested to determine which regression method works best on the data. The used parameters are described in Section 4.5.3. As shown in Table 4.3, Random Forest shows the best results across all feature sets. The number of input eigenvalues makes a negligible difference for the overall performance, but results in much faster training, as the dimensionality of the input vectors vastly decreases. One concern is the dimensionality of the input vectors versus the number of samples. While 30 eigenvalues per feature would result in a dimensionality of 180 for 469 training samples, we should keep in mind that the models are foremost training on the distribution of the top eigenvalues. As such, reducing the number even smaller results in only slight changes of the actual accuracy, as long as the top- n eigenvalues containing the actual characteristics of the similarity matrix are preserved. Sorted by size, for most concrete terms with very similar images, only the very first eigenvalues contain much information, with a long tail of close-to-zero values. For more noisy datasets of abstract terms, this might vary, so $n = 30$ was chosen conservatively to be on the secure side.

Table 4.4: Feature comparison for abstract words vs. concrete words.

Features		Abstract		Concrete	
		Correlation (1 : best)	MAE (0 : best)	Correlation (1 : best)	MAE (0 : best)
Low-level	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-level	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
Combined	L* (Low-level)	0.32	10.90	0.16	11.37
	H* (High-level)	0.27	11.31	0.10	9.10
	All (Proposed)	0.26	10.79	0.17	10.11
Comparative	Text [8]	0.40	13.27	0.18	7.51

In the fourth experiment, the effect on different visual features on the imageability estimation for high- and low-imageability has been analyzed separately. As words with high-imageability scores and words with low-imageability scores correlate with *concrete* words and *abstract* words, respectively, the visual characteristics of images in each word’s dataset are very different. While words with high-imageability scores share similar concrete objects or scenes, these with low-imageability scores have much more noise and mostly share similar *atmosphere*, *backgrounds*, or the like. When splitting the testing dataset into two parts around the median imageability score of the ground-truth scores, the resulting dataset can be classified as one half of *abstract*, low-imageability words vs. one half of *concrete*, high-imageability words. An analysis on what effect each visual feature has on the results of these subsets is shown in Table 4.4. The low-level features work better for abstract words, while the high-level features work better for concrete words. This shows that the visual features can in fact complement each other for different imageability words. The results also demonstrate that the concrete words have a lower average error of 9.10 than the abstract sub-sets with an error of 10.90. This is intuitive, as less imageable words are harder to grasp, as they do not create a clearly defined mental image (like *peaceful*), or are outliers which most likely create no reasonable dataset (conjunctions like *because* or *somehow*.)

Table 4.5: Feature comparison for different parts-of-speech.

	Noun (32)		Adj. (33)		Adv. (18)		Verb (11)		Misc. (24)	
	Corr.	MAE	Corr.	MAE	Corr.	MAE	Corr.	MAE	Corr.	MAE
L1	0.31	11.38	0.64	14.45	0.32	31.51	0.85	19.07	0.20	33.17
L2	0.35	11.02	0.27	18.32	0.14	33.35	0.90	20.35	0.27	31.45
L3	0.40	11.15	0.36	17.28	-0.02	32.33	0.89	20.27	0.43	29.57
H1	0.67	8.69	0.50	16.31	0.56	29.11	0.85	17.71	0.85	30.99
H2	0.38	11.07	0.23	17.07	0.35	36.46	0.81	22.44	0.00	36.12
H3	0.35	11.28	0.36	17.13	-0.10	37.83	0.56	25.91	0.28	34.06
L*	0.42	10.36	0.65	14.68	0.02	31.59	0.77	19.90	0.32	31.40
H*	0.60	9.05	0.47	16.36	0.51	28.31	0.76	17.81	0.49	30.78
All	0.65	9.17	0.53	15.42	0.29	29.08	0.79	18.13	0.60	30.47
Text	0.70	10.36	0.74	13.63	0.25	34.81	0.63	22.69	0.39	33.25

The fifth experiment shows preliminary results for different parts-of-speech. Similar to the analysis of abstract vs. concrete words, I was interested in how the performance of different features varies for different parts-of-speech. Unfortunately, the obtained dataset predominantly consists of nouns, resulting in too few non-noun samples for the random training-testing data split used in other evaluations. As a workaround, the regressor is trained with only noun-samples. This leaves all non-noun words for the testing dataset, which is sufficient to evaluate the trends for each part-of-speech. The results in Table 4.5 show that different features can excel for different parts-of-speech. Both the combined feature set using only the high-level features, and the one using the proposed combination of features can predict the imageability sufficiently across the majority of parts-of-speech. Similar to the overall results shown in Table 4.1, the high-level feature H1 shows the best performance as a single feature. As the model for Table 4.5 is trained using only *nouns*, it is no surprise, that the *nouns* have the smallest error. The hardest parts-of-speech to predict are *adverbs* and *other*, the latter one containing very non-visual terms like stop-words, conjunctions, and prepositions.

An example of some actual outputs of the proposed method is shown in Table 4.6, which compares the ground-truth annotations to the predicted scores for a selection of words. The three sections show words from the testing dataset, analyzing

Table 4.6: Imageability prediction results of the proposed method.

	Word	Predicted score	Ground-truth score
High-imageability	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-imageability	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 (Worst 5)	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

the results for words with high-imageability scores, words with low-imageability scores, and some outliers where the prediction failed, respectively. The examples show a close resemblance to the ground-truth scores, successfully predicting between Likert-scale levels of accuracy. The worst five outliers can show, that even in a wrongly predicted case, rounding to the next closest level in the Likert scale is usually at most by one or two scales off, preserving the general trend for most words.

To get a better understanding of the correlation between ground-truth scores and the predicted scores, Fig. 4.4 shows a scatter plot of the predicted testing dataset. Comparing the results of the proposed method with the comparative method, the global trend of each match very closely, but shifted along the vertical axis. Lowly-imageable words are such words that are considered to be harder to estimate due to their vagueness and abstractness. The scatter plot suggests that the proposed method works better towards words with lowly-imageability scores, despite the bias of the training dataset, compared to the text data-mining method from [8]. Note that while the proposed method only uses 469 samples for training, the datasets used in [8] were in average about a magnitude larger.

4.6 Discussion

In the previous sections, a method to estimate imageability using visual features has been proposed and analyzed. In the following, the results shown in Section 4.5 are discussed, including the implications that visual feature selection might have for applications using imageability and multiple modalities.

4.6.1 Performance and feature selection

In the best feature selection, the proposed method yields an MAE of 10.14 with a correlation of 0.63. Note that the error is relative to a regression to a range of [0, 100] for understandability of the results. As most Psycholinguistic based ratings are often expressed in Lickert scale, the results in Table 4.6 are converted to the range of [1.00, 7.00] to match the ground-truth scores. As shown, the error is smaller than one level on the Lickert scale (approximately 0.71 level), meaning that in average it successfully predicts the correct level of imageability. The number of evaluated words also ensures that the method is stable for a high variety of words. This means, it can be used as a tool to expand imageability dictionaries in an automated manner using image crawling and data-mining. In contrast, the previous work has only been evaluated on a small number of nouns within the same domain, and thus yielded a much higher error on the much higher scale of this dataset, including words across various domains and topics.

When evaluating the feature selection for different sub-groups of test data, the experiments led to interesting results. The error for abstract words is consistently higher than that for concrete words. This is not surprising, as abstract words are much more vague by nature, and thus are commonly harder to grasp, even for humans. The low-level features ought to capture characteristics as seen by the machine, while high-level features encode characteristics as seen by the human. Initially I was expecting that this would directly correlate to the performance for words with high-imageability scores vs. words with low-imageability scores. While single features

show mixed results on this, the combined feature sets using only low-level features or only high-level features confirm this assumption. The low-level features work better for predicting abstract terms, as they capture global concepts of the pictures, including atmosphere and mood. In contrast, the high-level features work better for concrete terms, which are often actual objects within each image. Looking at the information actually encoded within each visual feature, we can infer why they excel for different categories of words, as follows:

The color feature captures the *atmosphere* of the imageset. Even if the images otherwise show few visual resemblance, this feature can capture common *warm* or *cold* colors, for example. Additionally, abstract terms can often include technical figures or diagrams, containing lots of white background. In this way, color turns out to be a good choice for very abstract terms, where other visual feature can not find much similarity. The image theme and content features encode *actual objects* in the images. This makes them candidates for high-imageability words, as they are often connected to concrete objects and many images share similar objects.

When comparing the correlation results, it is noteworthy that there is a high correlation in the overall results shown in Table 4.1, and comparatively lower correlation when evaluating only abstract or only concrete words (as shown in Table 4.4). This indicates that the general trend of high- vs. low-imageability scores can be predicted successfully, but the order of words within each group is harder to predict. This is due to the limitations of the seven-level Lickert scale of the ground-truth scores. When looking at the dataset, many concrete terms are clustered closely around the level 6, while most abstract annotations are clustered around 3. Therefore, a small prediction error can reduce the correlation of close-by words, while the overall general trend is preserved.

Analyzing parts-of-speech, it is noteworthy that the words in each category show rather mixed characteristics. While intuitively, adjectives and adverbs seem highly imageable, as they increase information and context, they are often hard to put in visual context. For example, the word *red* can be directly expressed with visual

features (most prominently, the Color feature,) while words like *good* can not be matched to certain visual characteristics. The results also show, that some categories have a higher error than others. The category *other* contains words like *because* and *however*, whose datasets result in mostly random images. It is also noteworthy that the dataset predominantly consists of nouns, and thus the model was trained on only nouns.

4.6.2 Comparison to text-based methods

When comparing to Ljubešić et al.'s method [8], the evaluation shows that both text-only and image-only approaches can have different strengths. For the overall results, the proposed method using only visual analyses has a better MAE, while the textual approach by Ljubešić et al. [8] has a better correlation. This suggests that the predicted scores of the proposed method are closer to their ground truth, while the correct order might have some flipped results. On the other hand, the textual analysis has most results in a *more correct order*, while the actual error of outliers might be higher. This is especially true for the experiment splitting abstract and concrete words. Due to the nature of imageability being on a seven-level Likert scale, closely imageable words are very hard to rank in order, even for a human. On top of that, the dataset is biased towards the concrete end with the testing dataset having an average score of 70 of 100. As such, I believe that a correct order is vague and the general trend of the predicted scores is more important for many applications. Note, though, that it might heavily depend on the application, whether the correlation or the MAE is the better metric.

Another interesting result is that the textual analysis is better for concrete terms, while the visual analysis yields better results for words with low and mid-imageability scores. These results are probably also strengthened by the proposed method intrinsically focusing on noise analysis, and words with a high visual variety are usually highly abstract.

In terms of computational complexity, the proposed method using visual features took in the order of magnitude of several weeks for processing 5,000 image per word for 586 words. For this, the feature extraction was the major bottleneck. Note that, due to it being a pre-processing step only performed once, it was not further optimized. In contrast, the histogram comparisons and training took in the order of magnitude of a few hours for the full evaluation. Due to the results not being time-critical, there were no further evaluations or optimizations made for these either.

The text-only approach proposed by Ljubešić et al. [8] was not trained by ourselves, so it is hard to compare the computational complexity directly. Their paper does not comment on the computational complexity of their approach either. However, due to the nature of image vs. text processing, we can assume that a text mining approach would be slightly faster computational-wise. On the other hand, the evaluations showed that the visual analysis has advantages for certain words. Especially for more abstract terms, the scatter plot as well as the MAE show some advantages for the visual approach, while the textual approach can usually yield better correlation. For more concrete terms, surprisingly, the opposite is true. Therefore, a visual data mining in addition to a textual analysis can be an effective way to improve the accuracy of the imageability estimation.

4.6.3 Dataset

The results show that increasing the number of images for each word increases the performance. This seems intuitive, as more images equal to more data to be mined, and thus potentially more retrievable information. An increased number of images can also make the results more robust to noise. As far as complexity goes, the visual feature extraction scales linearly with the number of images, while the similarity matrix and histogram comparisons have quadratic complexity. The dimensionality of the visual features as well as the number of training samples have only major impact when choosing a Neural Network for regression, as the impact is negligible for the other methods.

Keeping this in mind, research by Sun et al. [15] suggests that there is no upper limit for improving machine-learned models by increasing the amount of data, but just a logarithmic diminishing return. Therefore, and due to the increased processing time, I have not further increased the number of images, although it can be assumed that the error can be decreased by it.

The number of words, on the other hand, seems to be sufficient to ensure stability within the prediction. Experiments with changing the number of training samples led to roughly similar results, which suggests that the number of data is sufficient to yield stable prediction. Note that the experiments were performed in the order of crawling, as more and more words became available with sufficient number of images in their imageset. This, however, means that the dataset with more samples also would include images for *harder-to-crawl* words, which could potentially decrease the performance through noise or word difficulty.

Lastly, I will summarize a few limitations and potential issues of the dataset creation process presented in this research topic. The switch from a recomposed custom dataset in Research Topic 1 to a direct crawling of crowd-sourced data had a variety of advantages, and makes for a vastly increased number of both words and images to evaluate. However, as a downside, the resulting dataset can become more noisy and potentially much more biased. As Flickr, in essence, is a Website for professional photographers, the images can be biased towards things photographers see as art, not fully capturing a neutral view on the concepts.

Looking at the outliers presented in Table 4.5, it also shows some points where using Flickr image for the results might not fit the expectation. Words like *fauna* result in many images in jungles, zoos, or similar backgrounds appealing to photographers. As such, they are visually rather similar, resulting in a high imageability prediction. The ground-truth score, however, is rather abstract, as the term is usually associated with biology, making it a rather *hard* and *scientific* word. In contrast, words like *email* or *plastic* result in rather noisy datasets, as it is not really clear, what kind of photos people would upload, tagged with these words. As a result, the prediction for

both is mildly imageable. For the ground-truth score, however, these are considered highly imageable, mostly because they are considered to be objects, or rather, in case of e-mail, with a concrete *thing* people often deal with.

Another downside is that it is hard to obtain single images for parts-of-speech like conjunctions, verbs, and stop-words. The nature of these types of words unavoidably results in the image data of these words to be random images or non-related. Note that many conjunctions and stop-words are naturally rather abstract and lowly imageable, so the data-mining will potentially still lead to good results for these terms, *especially because of its random nature*. Similarly, the current method makes no difference between ambiguous meanings. As such, the imageset for *craft* might be a mixture of *handcraft*, *aircraft*, and *watercraft* (which arguably makes the term rather abstract due to the ambiguity.)

4.7 Summary

Chapter 1 introduced the concept of quantifying the mental image using image data. The problem was divided into the two tasks of estimating relative and absolute measurements. In this chapter, the second sub-task of estimating absolute measurements has been tackled by means of an algorithm-driven method proposed for absolute visual variety measurements in form of imageability score estimation.

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

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 highly imageable versus lowly imageable is sufficiently 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 in multimodal applications like sentiment analyses. Another possible application which comes to mind is 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.

Chapter 5

Analysis on relative and absolute approaches to visual variety

Chapter 1 discussed the motivation of this doctoral research, dividing the problem of mental image quantification into two sub-problems; relative and absolute measurements. Following, previous two chapters 3 and 4, first proposed a method to estimate the relative visual variety of concepts in a narrow domain (Research Topic 1), and then a method to estimate absolute imageability measurements for arbitrary concepts on the dictionary-level (Research Topic 2). In the following, some overall comparison and discussion regarding both topics are outlined. Section 5.1 will comment on the core assumption stated in Chapter 1 regarding the relationship of the mental image across society and images crawled on the Web. Next, in Section 5.2, the relative estimation and absolute estimation are compared, especially regarding their applications. As both methods chose different approaches to solve their sub-problems, Section 5.3 will discuss some advantages and downsides of data-driven methods vs. algorithm-driven methods for this use case, as found through the experimental results discussed before. Section 5.4 will give some comments on the reproducibility of the proposed research topics.

5.1 Core assumptions

In Section 1.3, the idea of using Web-crawled data for visual perception modeling was introduced. 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 (Fig. 1.5(b)). 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.

The core assumption discusses the quantization of *an average mental image across society*. Section 5.1.1 discusses the problem with personalized scores. There is some caveat to using Web-crawled data; using Web-crawled data means one needs to deal with noise and biased data. Section 5.1.2 outlines biases found when preparing and analyzing the datasets used for the proposed methods.

5.1.1 Personalization

One thing that the contributed methods did not consider is personalized scores. This is due to the core assumption discussed before focusing on the average mental image across society. In Research Topic 1, the dataset was composed by considering Web popularity of sub-ordinate concepts. This approach assumes a popularity score with which most people would agree with. If future work were to adjust this method for personalization, a personalized popularity metric could be the first step. In Research Topic 2, images were crawled from Social Media sources, intrinsically already assuming crawled data having a composition most humans would perceive reasonable. Thus, this approach can not easily be adjusted for personalization. Meanwhile, the data-driven nature of Research Topic 1 would allow more options towards personalization.

5.1.2 Dataset bias

When dealing with images crawled from the Web, one needs to deal with noise and bias.

Regarding noise, the core assumption intrinsically assumes noise to exist. For more abstract words, the mental image is less clearly defined, resulting in a higher variety, but also with higher noise. Words like *something* are very vague, but would also result in a very high ratio of noise, as it is also not clear, which image it would result in. In contrast, a crawling for *airplane* would probably result in only minor noise, if any. As the proposed methods are built upon the similarity of visual feature spaces and thus associating a high variety with high abstractness, noise would intrinsically *help the approach to work*.

Bias, on the other hand, is something that might be problematic. When dealing with Web-crawled images, I intrinsically assumed that concepts with many images available would also be the ones of the most interest. We could argue that *most common* and *most popular* are different things. Following, the popularity of *sports car* used in Research Topic 1 resulted in very high measurements for almost every metric.

An interesting outlier I have found in the analysis is the word *canon*. WordNet would associate this word with a piece of music, a collection of books, or a body of rules. However, many images found on Social Media are tagged with its camera model, and thus *canon* resulting in any image shot with a certain camera brand. In this case, it is arguable which meaning of the word is actually more prominent (e.g., the camera brand or the terminologies cited above.) The proper noun *canon* (the camera brand) has more significance because of Flickr being a photographers-targeted service biasing the results. There is also a polysemy-related issue, as it is hard to determine which meaning to crawl when dealing with text-based queries for image accumulation. As a result, homonyms can result in noise, as the proposed methods can not distinguish between the different meanings. Therefore, the query of *craft* could result in images of either *aircraft* or *handcraft*. While this is a prime

example of the semantic gap on its own, this problem resulted in biased approximate datasets, as it was often unpredictable which kind of images it would result in.

One way to decrease noise regarding homonymy or polysemy could be filtering based on word embeddings. By relating the meaning of a concept to the surrounding textual metadata of crawled images, one could filter images that are unrelated to a given word meaning. In Research Topic 1, each concept is based on a synset, so this approach would potentially work. For Research Topic 2, the imageability dictionaries do not have a description attached. Therefore, it would not be possible to clearly determine a single concept in case of homonymy or polysemy. Note that the approaches intrinsically assume a high noise ratio for more abstract terms, so filtering might negatively affect the results.

5.2 Relative measurement vs. absolute measurement

Two different approaches for mental image quantization were discussed. One approach targeted relative measurements (Research Topic 1) and the other absolute measurements (Research Topic 2). It is considered that neither approach is necessarily *better*, but it rather depends on the target application, which one results in more favorable measurements. In the following, applications for relative measurements and absolute measurements are discussed in Section 5.2.1 and Section 5.2.2, respectively.

5.2.1 Applications for relative measurements

The target of Research Topic 1 was the relative comparison of concepts in a limited domain. The experiment focused on the domain of **vehicles**, comparing concepts such as **cars**, **trucks**, or **airplanes**. The approach uses a common reference point, **vehicles**, which is considered the most abstract. As such, all other concepts are sub-ordinate and contain less variety. This is crucial for the estimation, as a too unrelated reference point would yield unusable results. The group of vehicles, however, is semantically well connected and thus results in repeating image backgrounds, situations the depicted vehicles are in, and so on. The proposed idea is considered to work similarly for other narrow domains like **animals**, **plants**, and so on, but probably not for too unrelated words, like the hierarchy of all **concepts**.

One of the main target applications I had in mind when approaching this research idea was image captioning and image tagging. For image tagging, the choice between many possible candidates occurs frequently: An image of a car could be tagged **vehicle**, **sports car**, or **car**. All of them would be too verbose, but there are no existing metrics for quantitatively comparing the candidates. For such an area of word selection problems, Research Method 1 seems best fit. As a relative measurement, it can find minute differences where a dictionary-level comparison would have too cluttered results to find a meaningful ranking.

In the field of *Explainable AI* (Section 2.4.2), the goal is to bring light into black-boxed approaches in machine learning and artificial intelligence. With increasingly more convoluted approaches to multimedia processing, a need for a better understanding of vision and language becomes important. Similarly, as discussed in the work by Hentschel and Sack [70], a machine might not necessarily see the same as a human. 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, this research could be considered as an assistance to these and related semantic problems.

Here, the results of the proposed method would be of interest, revealing hidden semantics in imagesets, a human might not notice. As the results of Research Method 1 are not trained on ground truth but are rather simply a comparison of clustering through conventional methods, it has the potential to be less biased than a method relying on training.

5.2.2 Applications for dictionary-level absolute measurements

The target of Research Topic 2 was the absolute comparison of arbitrary concepts. As such, the possible target space would be the whole English dictionary, comparing concepts like cars with pizza or peaceful. To associate the visual characteristics with semantic knowledge, an imageability dictionary is used for training. While this makes it easier to obtain estimates for arbitrary words with a single trained model, it also results in a much more cluttered scale of outputs.

In multimodal applications using text and images, concrete details and abstract concepts often supplement each other. Imageability as a concept could be used to assess the quality of auto-generated captions for a given application. Due to its nature, a caption to be used in an image retrieval application does need different contents than those of an image caption in a newspaper article. The imageability of words used in the caption gives an indication on how descriptive it is, and whether the reader would

be able to easily mentally visualize it. It could be used to assess the accessibility, or the degree of information, in auto-generated texts.

The absolute nature of the proposed model means that a single model can be used for arbitrary words. Following, extending imageability dictionaries is facilitated and can be done for any unannotated word in reasonable time. This makes the approach feasible for multimodal approaches with a large number of unannotated words, like image captioning [12]. As the method relies on analyzing a set of images, it would also be possible to create datasets for proper-nouns, like names or places, for which by nature no entry in imageability dictionaries exist. Furthermore, the results could be included in tools and datasets for NLP and sentiment research, like Empath [63], as they provide additional insight on semantic text understanding.

5.3 Data-driven vs. algorithm-driven approaches

In the previous chapters, ways to solve the mental image quantification for absolute and relative comparison were discussed. For each research topic, different approaches were applied to solve the sub-problem. In Research Topic 1, a data-driven approach was chosen, where existing data is reconfigured to resemble the human perception of dataset composition. This was considered to be promising for the narrow target domain and proved promising for the relative measurements, but it came with some issues. In contrast, in Research Topic 2, an algorithm-driven approach was chosen to solve the absolute approach to mental image quantification, which could work on arbitrary concepts. As each approach come with its downsides, they are briefly discussed in Section 5.3.1 and Section 5.3.2, respectively.

5.3.1 Problems of a data-driven approach

In Research Topic 1, the task of measuring the relative visual variety of concepts in a narrow domain is tackled using a data-driven approach. The core assumption is 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 (*car*) may have a stronger influence than a concept rarely seen (*jet*). Thus, for each concept, a custom imageset is created using the WordNet [45] hierarchy of its hypernyms and a popularity measurement to determine the importance of sub-concept images in its parent imagesets. The resulting imageset is considered to be *ideal*, meaning that its composition resembles the frequency of subordinate concepts in real life, which is assumed to directly relate to the visual variety of the parent concept. While the approach led to promising results, it comes with several downsides:

First, the number of images available for very obscure sub-concepts could heavily bottleneck the recomposition of its parent concepts. This was especially true, if the popularity of the said sub-concept was estimated unexpectedly high, be it through noise or simple error.

Second, as it is tied to WordNet [45] and ImageNet [25], it would not work for words which are not available in both datasets. ImageNet is only available for nouns and is rather limited in the number of terms available. It also fully relies on a hierarchy of hypernyms and hyponyms, which are not available for anything but nouns. Adjectives, for example, would lack both a baseline dataset as well as a hierarchy used for recombination.

Third, a proprietary API was used to estimate a popularity metric for sub-concepts based on Web search engine hit results. This led to unnecessary cost, and semi-manual automation of scripts to reconfigure the datasets.

Lastly, the process mainly concerned the recombination of parent concepts, so the leaf node concepts would not benefit from the majority of the proposed contributions. Due to these limitations, its evaluations could only be performed on a rather limited dataset of 25 terms related to vehicles, and about 2,400 images each.

5.3.2 Problems of an algorithm-driven approach

In Research Topic 2, the task of measuring the absolute visual variety of arbitrary concepts on a dictionary-level was tackled using an algorithm-driven approach. Due to its nature, a purely algorithm-driven approach can not deal with dataset bias. Furthermore, in its current state, the algorithm-driven approach would always clutter similarly composed imagesets together. Allowing the comparison for arbitrary concepts means focusing on the overall trend —e.g., is a concept very imageable or very unimageable. The correlation results show that it is very hard to maintain a correct order, especially when looking at close scores. Comparing only concrete words for example, the results will be too cluttered to find minute details. Here, the hierarchical data-driven approach will have its clear advantages, as the composite nature of different datasets will always force a clearer ranking of related concepts.

5.4 Reproducibility of published work

In recent years, the reproducibility of academic results have become more and more focused on in the research community. Even more important is the availability of source codes, as this means other researchers can directly build their ideas on top of existing methods, rather than needing to reimplement them on their own.

For Research Topic 1, being a data-driven method, the actual implementation is a combination of intermediate results for popularity-based metrics, image sets from various sources (ImageNet, supplemental images from Google [37] and Bing [17]), and some scripting to automate the recombination of each imageset. Dealing with multiple APIs and existing datasets, publicizing the source codes does not seem too meaningful. Adjusting the scripts for local file structures would be almost equivalent to rewriting them. Following, the actual source codes have not been made available yet.

One source of concern is the black-boxed nature of the APIs for dataset retrieval. While the actual crawled images can not be redistributed for copyright reasons, 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 could be that it can adjust to changes over time, as recrawling the data also *updates* it regarding to how popularity might have changed over time.

For Research Topic 2, the source code has been made available on GitHub ¹. Additionally, two students in the lab have been using this framework for their own researches. While doing so, they reproduced the results multiple times with both the datasets used in the experiments of this thesis, as well as datasets crawled from other sources like Google [37] and Bing [17]. This strengthens the proposed method as

¹<https://github.com/mkasu/imageabilityestimation/>

well as the main assumption, as it means it not only works on my own data, but with any reasonably Web-crawled data.

The original dataset is crawled from YFCC100M. While this dataset can not be re-distributed due to copyright concerns, it is a public dataset that can be recrawled in exactly the same way as it has been done for the experiments in this thesis. The imageability annotations are part of existing imageability dictionaries, so they are also publically available.

Chapter 6

Conclusion

This chapter concludes this doctoral thesis. In Section 6.1, the proposed contributions of this doctoral research are summarized. Section 6.2 discusses remaining challenges in this field of research and potential directions for future research, both towards extending proposed methods as well as applications for the proposed metrics of visual variety and imageability. Lastly, Section 6.3 completes the thesis with closing remarks.

6.1 Summary

The research described in this thesis attempts to quantify the perceived variety of concepts from a visual standpoint. Although the semantic gap describing the lack of coincidence between computer representations and human expectations has been a core issue in the field of the Multimedia research field, there is a lack of understanding of semantic distances between abstract and concrete concepts. Following, multimodal research involving abstract concepts is an ongoing challenge. This results in word choice problems in image captioning or machine translations, among other problems. With this semantic gap between vision and language, this doctoral research aimed to find a measurement of perceived differences between concepts.

As a core assumption, images crawled from Web and Social Media were assumed to intrinsically contain knowledge on the average perceived mental image through their dataset composition and image contents. As such, an analysis of the visual feature space of two concepts' datasets would yield knowledge on how these concepts are perceived differently. Following, the goal was to compare image datasets for different concepts to quantify the perceived variety differences of those datasets.

To tackle this idea and verify the core assumption, the main problem of mental image quantification was divided into two sub-tasks, which were proposed and tested: Relative and absolute measurements. First, in order to measure the relative visual variety of concepts in a narrow domain, datasets for related concepts were composed based on their sub-ordinate concepts and then compared in a data-driven approach. Next, in order to measure the absolute visual variety of arbitrary concepts, images crawled for general-purpose words were crawled from Social Media and analyzed in a variety of visual characteristics, in order to train a model to estimate imageability scores.

The first research topic discussed this idea in Chapter 3 with a data-driven approach to analyze the relative visual variety differences of a limited domain of related concepts. The proposed method creates a recomposed dataset for each concept based on their sub-ordinate concepts. Using a weighting, the ratio of sub-ordinate concepts can be changed, and thus how they influence the overall dataset composition and overall image of the dataset. For the experiments, different compositions were created using weightings from Web-based APIs. The datasets for each concept are then analyzed, and their visual feature spaces are clustered to obtain a variety measurement for each concept. The proposed method was tested using 25 concepts related to the domain of vehicles. It was evaluated using ground-truth scores obtained through crowd-sourcing, where people were asked to judge the perceived variety of two concepts in paired comparisons. Each dataset was compared to an unmodified baseline dataset, finding that the recomposed datasets result in a more natural clustering. Following, the created datasets more closely resembled the ground-truth visual variety scores obtained through crowd-sourcing. The proposed method yields

a Mean Squared Error (MSE) of 9.01 and a correlation of 0.73 for the selection of words related to vehicles. These results could be used to improve ontologies, as an evaluation metric for word choice problems, or similar problems looking at relative differences of concepts.

The second research topic discussed this idea in Chapter 4 using an algorithmic approach to compare a variety of visual characteristics across datasets to estimate absolute imageability scores of words for arbitrary concepts on a dictionary-level. Targeting the concept of imageability coming from the field of Psycholinguistics, this research topic applies the aforementioned core assumption to existing Psycholinguistic ground-truth scores describing the perception of words using a Lickert scale. The approach analyzes the visual feature space of Web-crawled images for words across a selection of six low- and high-level visual features to create cross-similarity matrices for each word. Using the cross-similarity matrices, a model is trained to regress the imageability score for the input word. In the experiments, 586 words and 5,000 images each were used to evaluate the method. The proposed method predicted absolute imageability scores with a Mean Absolute Error (MAE) of 10.14 for scores normalized to the range of [0,100] and a correlation of 0.63. This indicates a prediction of an average error less than one level on the Lickert scale used in the ground-truth Psycholinguistics research. The extended analyses also showed that combining low-level traditional computer vision features with higher-level neural network-based features tends to complement each other. Either set of features works better for a different sub-group of words, with low-level features like color histograms excelling for abstract words and high-level features from YOLO [65] excelling for concrete words. These results could be used for extended existing psycholinguistic dictionaries in an automated, non-labor intensive manner.

The two aforementioned 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. 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.

Analyzing the perceived semantic gap from two directions, applications can also combine both ideas. The first research topic looks at the semantic gap as a relative measurement in order to get a better understanding of related concepts, while the second research topic, in contrast, looks at an absolute measurement in order to find a general trend of the perceived gap even for unrelated concepts. As these concepts are tangential, multimodal applications can profit from using a combination of both types of measurement. For example, one could first get an understanding of the overall trend of unrelated concepts with an absolute measurement. In a second step, more fine-granular word choice problems could be tackled using the relative measurement. As such, a combination of both gives a comprehensive idea on the human perceived semantic gap between vision and language, valuable for many applications. This solves the aim of this doctoral research introduced in Chapter 1.

6.2 Remaining challenges and future directions

This thesis described the quantization of the perceived variety of concepts from a visual standpoint using two approaches. In the following, some general research directions for the estimation of such metrics, are discussed and some remarks on future steps for the individual research topics are given. In the end, opportunities for applications built using the proposed metrics are briefly discussed.

One thinkable direction of this research is a context-aware measurement of visual variety or imageability. The core assumption stated in Chapter 1 emphasized the aggregated nature of the Web-crawled images, and thus the goal for this thesis was the quantization of an average mental image across society. While this problem is difficult on its own, an even more difficult problem would be the consideration of the context for the estimated scores. For example, in the current state, the methods do not consider profession- or culture-based differences regarding the perception of words.

Research Topic 1: In this research topic, recomposed datasets for related concepts are created based on their sub-ordinate concepts. The approach has several bottlenecks in labor and cost through the way it relies on Web APIs for crawling images and estimating the popularity of sub-ordinate concepts. For example, the popularity score for certain words might be biased, sometimes creating a bottleneck during the recomposition step because of an insufficient number of images for a certain sub-ordinate concept. With a more elaborate way to handle noise or bias, the dataset recomposition could be automated, also allowing larger scale experiments. Combining the results of multiple APIs would be a first step, presumably resulting in fewer outliers, and thus, fewer bottlenecks. It would be interesting to use the measured results to improve ontologies.

While the recomposition of datasets uses a Web-based popularity measurement, a combination of different measurements might improve the results further while keeping the research less dependent on the results of a single Web API. A more sophisticated weighting of sub-ordinate concepts during dataset recomposition might be a good starting point for context-aware measurements. With this, the dataset composition could be changed based on the individual expected composition for different professions or cultures.

Research Topic 2: In this research topic, the idea of visual variety is applied to the concept of imageability from the field of Psycholinguistics, estimating the imageability of new words using Web-crawled images. The results verified the method as accurate using only visual data analysis. Interestingly, parallel research using nothing but textual data shows a similar accuracy [8]. Following, it would be interesting to combine various sources of knowledge, like textual and language information with the proposed method using visual information.

In previous research as well as the contributed methods, imageability has been defined as a concept describing the perception of a single concept. For nouns, this idea is clear forward, but when looking at future multimodal applications, one needs to consider measurements for other types of part-of-speech. The results indicated a still improvable performance for adjectives, adverbs and verbs. One way to consider these would be looking at phrases rather than single concepts. For an adjective, it would be intuitive that the imageability of such words is relative to the imageability of the nouns and verbs they syntactically modify. As such, a phrase-based or sentence-based definition of imageability scores might be one direction of future research.

Another direction for future research is the use of visual variety measurements as a source of knowledge for other multimodal applications. Existing research shows the viability of psycholinguistic features for comparing the meaning of cooccurrence of text and images [31]. Similarly, the metric could be used to evaluate texts in terms of their abstractness or visualness. If training absolute measurements of imageability

on image-based and text-based sources in separate, one would be able to map the relationship of those in a mixed vector space. With this, the appropriateness of images and captions could be inferred for a given concept based on their imageability scores. It could be used to derive a method comparing image captions on their applicability for different use-cases, similar to affective captioning. In Appendix A, two applications that analyze characteristics of datasets used for the analyses are discussed in more detail. They are proposed as ideas looking into future opportunities for visual variety related research.

6.3 Closing remarks

In Chapter 1, the stated aim of this thesis was to quantify the perceived variety of concepts from a visual standpoint as a way to measure the semantic gap between vision and language. The core assumption was that images crawled from Web and Social Media intrinsically contain knowledge on the average perceived mental image through their dataset composition and image contents. This assumption was verified through their dataset composition and image contents. This assumption was verified through both research topics, applying to both relative and absolute measurements of perceived variety across sets of concepts. While recent state-of-the-art methodologies increasingly focused on deep-learned models, the results found in Research Topic 2 show that both traditional and neural network based methods contain complementary semantic knowledge, equally contributing to the results found.

In the past decades, computer vision and Natural Language Processing (NLP) were mostly considered separately, each working with their own set of tools and approaches. With the former field of research analyzing only visual data, and the latter only looking at textual data, either community often does not consider the connection of both. However, recent progress in multimodal applications proved the rising need for connecting both research fields. Understanding the connection of vision and language becomes more and more important, moving these communities closer together. This is increasingly reflected in Multimedia research for image retrieval or image captioning purposes.

The human, on the other hand, is still not in the focus of the considerations. Following, many multimodal applications result in unnatural results, as if a machine created the results. Human perception related tasks like imageability, memorability, visual interestingness, and the like are just the first step in the direction of getting multimedia applications more natural and more similar to what a human would create. Although this doctoral research may only be a first step into the direction of fully understanding the semantic gap of vision and language from the viewpoint of a human, hopefully, the proposed contributions can incite future discussions and future research in this direction.

Appendix A

Dataset visualizations

This appendix outlines two visualization projects built for analyzing datasets used in Research Topics 1 and 2. They also serve as proposed ideas for future research directions and applications.

Project A.1 uses per synset datasets like ImageNet [25] or the ones created through Research Topic 1. Using a Bag-of-Visual-Words (BoVW) model, the visual characteristics of related concepts are outlined in the form of highlighted feature maps. With an interactive UI, it is possible to browse related sub-ordinate concepts and compare the visual variety of different concepts by highlighting the most important visual characteristics of each.

Project A.2 uses a visual sentiment dataset as its baseline. For each image, the textual and visual relationship is analyzed by calculating per-image psycholinguistics scores. A spatial embedding visualizes textually related images close to one another. The method uses imageability, among other word ratings, to find similarly perceived images. This demonstration is a use case of psycholinguistic word ratings for multi-modal ratings. As such, it showcases a possible future direction of research if psycholinguistic dictionaries can be extended through methods like that proposed in Chapter 4.

A.1 Visualizing Bag-of-Visual-Words models across related concepts

In recent multimedia applications, approaches involving text, image, and video contents are often used to combine knowledge spanning multiple modalities. The so-called semantic gap describes a number of problems that occur when transferring between modalities. Visual semantics can give an insight into human perception of given concepts. For example, largely overlapping sub-concepts might be less distinguishable, if they are also visually equal. In contrast, two very related concepts are more easily distinguishable, if visually distinct, even if they logically belong together. In Psycholinguistics, these properties are called *imagability* and *concreteness* [78]. A quantification of this would greatly benefit word selection problems in various applications.

For this visualization, datasets composed of logically related concepts are visually analyzed. A dataset is created by combining images from ImageNet [25] using the WordNet hierarchy [45]. A separate Bag-of-Visual-Words (BoVW) model is trained for each concept, using images of all its subordinate concepts. The model will prioritize keypoints standing out when visually comparing different concepts. By visualizing the resulting feature space spatially, hidden visual semantics of logically-related sub-concepts are shown. To aid in finding hidden semantics of related concepts, the most common visual words of an image in relation to its neighbors are highlighted. This provides an additional semantic knowledge on how sub-ordinate concepts visually relate to each other, laying the ground work to estimate psycholinguistic ratings like *imagability* and *concreteness*.

Section A.1.1 gives a brief overview of related work. In Section A.1.2, the proposed idea is introduced. First, the creation of the dataset and the visual model are described in detail. Then, the approach to highlight important visual words, as seen by the machine, is outlined. Section A.1.3 showcases an interactive UI, discussing possible gains in semantic knowledge through it.

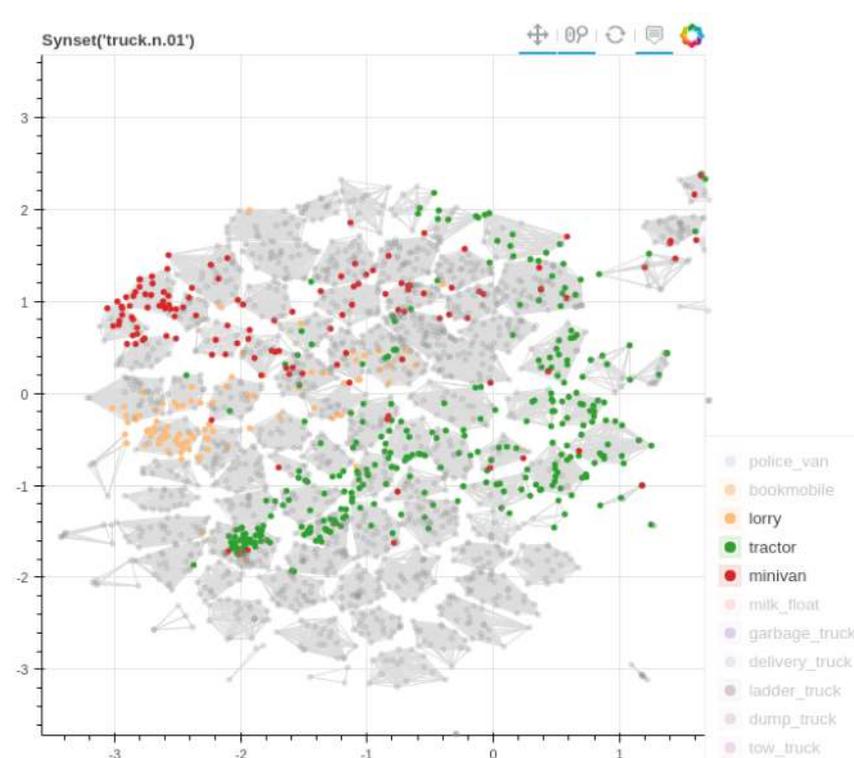


Figure A.1: Example of a visualized synset.

A.1.1 Related work

Research on how language interacts with human perception has been part of Psycholinguistics research. Paivio et al. [28] analyzed the concreteness, imagery, and meaningfulness of nouns. In the MRC Psycholinguistic Database by Wilson et al. [80], words are rated by familiarity, concreteness, imagability, and meaningfulness. More recent research by Cortese et al. [30] classifies imageability ratings for 3,000 words, which is thought to be useful for human word recognition and memory studies.

In Research Topic 1, the visual variety of concepts was quantified. The evaluation verified the estimate to match human expectations by comparing it to the results of a crowd-sourced annotation.

There has been research on visualizing visual feature spaces, as it is of interest to

understand how well object recognition works. Yue et al. [79] visualized the contents of a BoVW model. Using a modified model, they reverse the encoding and reconstruct images from the visual model. With this, they could visualize the degree of information lost in the representation. Hentschel et al. [70] use an object recognition classifier to visualize which regions of an image most likely contain the trained object. For a given image of an object, they create a probability heatmap highlighting which regions of the image most likely contain the object. Both projects use a feature visualization to judge the quality of the visual feature representations. However, there has not been any research analyzing the semantic implications of such visualizations.

A.1.2 Approach

Through the proposed visualization, the visual similarities within a group of related concepts are visualized. WordNet provides a hierarchy for every group of synonyms with a shared meaning, so-called *synsets*, using the hypernym/hyponym relationship of words. A synset is defined as *abstract*, if there are hyponyms in the hierarchy, and thus, if there are subordinate concepts which are classified below this concept. For every abstract synset, a dataset is created using images from subordinate concepts. A visual model based on a BoVW is computed for each abstract synset separately. Lastly, the most important visual words for each image considering their visually closest neighbors are computed and highlighted.

The goal is a visualization as shown in Fig. A.1.

A.1.2.1 Dataset

To analyze visual relationships within concepts, a dataset that has a strong variety of subordinate concept images is needed. For each abstract synset, a set of related sub-concepts is generated by crawling its most subordinate concepts in the WordNet hierarchy. The most subordinate concepts in the WordNet graph are the leaf nodes

below the abstract synset. This process is similar to the one illustrated in Figure 3.1 in the introduction of Chapter 3. Then, an imageset is generated using ImageNet images as a baseline. Instead of using the imagesets provided by ImageNet directly, the images of its sub-concepts are merged. This is intended to provide a dataset with a higher variety, and thus preserving knowledge about hidden concept semantics. The information on which image belongs to which sub-concept is preserved for labelling.

A.1.2.2 Visual representations

As a visual representation, a BoVW model is generated for each abstract synset separately. It is trained using images of its subordinate concepts using the previously created imagesets. For each image, Speeded Up Robust Features (SURF) [43] are used as visual features. SURF features are local gradient descriptors describing re-occurring visual patterns in the imagesets.

This model learns the visual differences of different subordinate concepts, as seen by the machine. Thus, the visual words will encode keypoints which stand out relative to other subordinate concepts.

A.1.2.3 Visualization

For visualization purposes, Uniform Manifold Approximation and Projection [74] (UMAP) is used to compute a dimensionality reduced spatial embedding of the visual model. This embedding gives insight into the spatial distribution of different subordinate concepts within the visual feature space. Next, the goal is to highlight the most common visual words, as seen by the machine, in relation to neighboring images. This allows inferring what the computer perceives as visually related parts of neighboring images. The process of selecting the most common visual words for each image is shown in Fig. A.2. For each image, a number of visually similar images are selected using Mean-Shift Clustering [16]. Then, the BoVW histograms

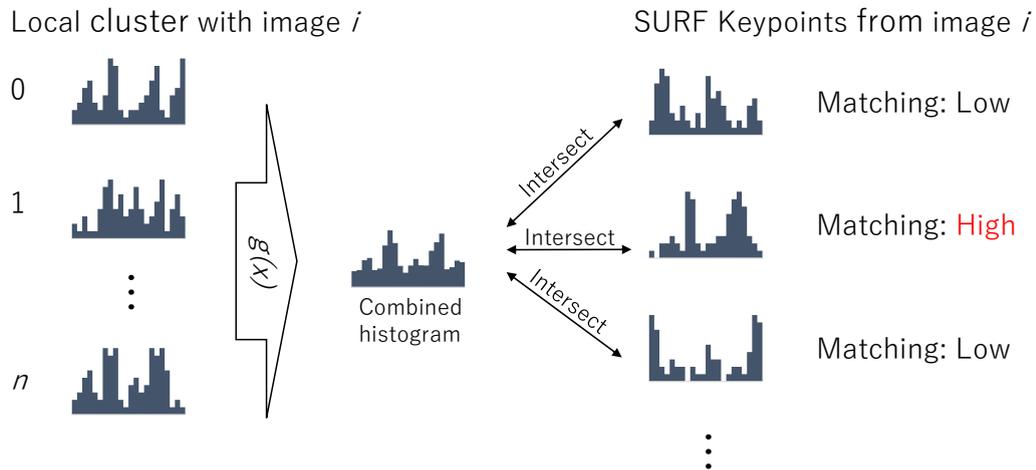


Figure A.2: Process of selecting common visual words.

$f(x)$ of all selected n images are merged using this equation:

$$g(x) = \frac{\prod_{i=0}^n (f_i(x) + 1)}{2^n}$$

This method will create a combined histogram $g(x)$ with amplified common peaks. It biases the distribution for the most common visual words. For each single keypoint of a given image, a BoVW histogram is computed and intersected with the histogram $g(x)$. The top 10 percent closest matching keypoints are selected as important regions for visualization.

Figure A.3 shows an example of four neighboring images in an imageset within the imageset for the synset truck. While they belong to different subordinate concepts, they share visual similarities and are thus clustered together. The red regions in the bottom row highlight the most common visual words. As all images are shot from a similar angle, features around the vehicle roof and front glass are the most common. Following, the machine understands the trucks visually to have a long rectangular shape, e.g., discerning it from a car, with some specific visual characteristics around the front window, e.g., making it a motor vehicle.



Figure A.3: Example of the common keypoint visualization for the synset “truck”.

A.1.3 Visualization tool

Using the visualization framework Bokeh [77], an interactive tool to visually inspect synsets has been developed. It opens a pre-processed synset, showing the spatial embedding of its visual feature space using UMAP.

Labels for subordinate concepts can be displayed to view the spatial distribution of those concepts within the visual space, as shown in Fig. A.1. It can highlight labels for the most-subordinate concepts (children), or display subordinate trees going from the root synset (siblings). The area where samples of a subordinate concept span, can give insight into the variety and abstractness of that concept. Furthermore, the overlap of image clusters can show how visually similar sub-concepts are seen by the machine.

When hovering data points, the original image and the BoVW visualization are displayed, as shown in Fig. A.4. This can be used to compare neighboring images and discover which visual characteristics are seen as useful for the machine when classifying these images.

A.1.4 Comparing image regions

If training the visual model on full images, features in the foreground and background are treated equally. For the use-case of evaluating semantics across different

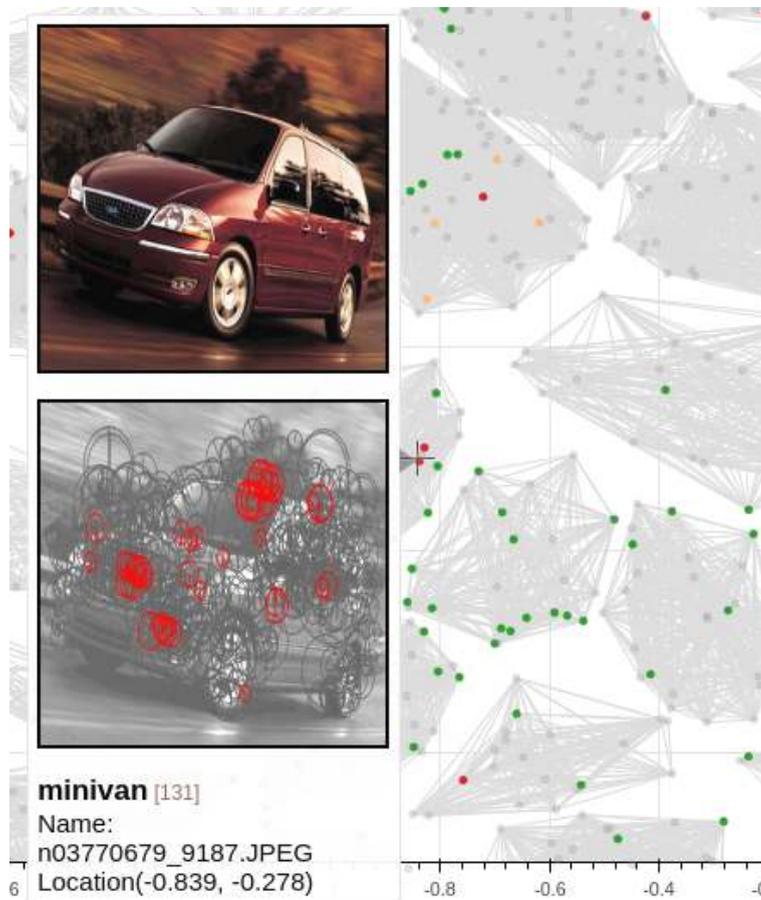


Figure A.4: Visualization tool shows detailed information as well as the keypoints when hovering the mouse over the datapoints.

concepts, this might actually be beneficial as the background includes extra semantic information, not otherwise available in the visual data. For example, the visualization can showcase clusters of concepts, where the background plays a more important role than the foreground. In a visual concept containing all *vehicles*, images of *helicopters* and *airplanes* might be clustered together, as the visual characteristics of clouds are visually more important to understand those images than characteristics along their chassis.

In an object recognition context, this will inevitably lead to unfavorable results, needing for image segmentation. When assessing the human, however, concepts might create a similar mental image, if they appear in similar situations. Therefore, an

analysis of common backgrounds might help in estimating properties like familiarity, concreteness, imagability, and meaningfulness.

A.1.5 Summary

A tool to visually compare logically related concepts has been introduced in this section. Using a spatial embedding of a BoVW model, visual characteristics like feature variety of related sub-concepts can be assessed. By amplifying common peaks in neighboring BoVW histograms, common visual words are extracted and highlighted. This showcases how the machine perceives visual differences of images, which can emphasize hidden semantic knowledge.

The visualization allows finding interesting similarities between neighboring images. Comparing the area spanned by subordinate concepts, the visual variety can be grasped. The tool can find perceptually indistinguishable sub-concepts by highlighting an overlap in their labels.

For future directions, one could look into including other visual features, especially neural network-based features. Furthermore, it would be interesting to include a heatmap based visualizations of visual feature importance in addition to the keypoint visualization.

A.2 Visualization of image sentiment datasets using psycholinguistic groundings

The use of text and imagery from Social Media for tasks related to sentiment and emotion research became ubiquitous in recent research. However, there has been little research regarding the multi-modal implications of images and their annotations related to human perception. In this section, a tool to visualize psycholinguistic groundings for a sentiment dataset is introduced. Using this, the relationship between texts and images, trying to get a better understanding of the groundings of human perception can be analyzed. For each image, individual psycholinguistic ratings are computed from the image's textual metadata. Combined with sentiment scores available from the used dataset, a sentiment-psycholinguistic spatial embedding is computed. It shows a distribution of sentiment images close to human perception. Based on this, an interactive browsing tool, which can visualize the data in various ways, has been created. The tool allows highlighting different psycholinguistic ratings in heatmaps separately, as well as to understand the structure of different datasets based on their ontology.

Section A.2.1 then discusses the idea of combining the sentiment scores of a given dataset with psycholinguistic groundings from the image metadata to compute individual scores for each image. Lastly, Section A.2.2 showcases the interactive dataset browser built to visualize embeddings of the sentiment-psycholinguistic space, which can be filtered across different nouns and adjectives. Various color modes allow for highlighting the different sentiment and psycholinguistic ratings.

A.2.1 Approach

Here, the aim is to present a means to analyze psycholinguistic groundings for sentiment image datasets. As a first step, a visual sentiment dataset having a large number of images annotated with Adjective-Noun Pairs (ANPs) is retrieved. Using

the textual metadata attached to an individual image, nine psycholinguistic scores are computed for each image. Lastly, a set of spatial embeddings based on each individual images' sentiment-psycholinguistic scores are computed for each noun, adjective and ANP, respectively.

A.2.1.1 MVSO dataset

The MVSO dataset [82] is used as the baseline for the visualization tool. The dataset consists of seven million images, their textual metadata, and sentiment scores, collected through Flickr and crowd-sourcing. Each image is annotated with a single ANP, e.g., *abandoned city* or *old dog*, describing its sentiment. The ANP is split into two labels; *noun* and *adjective*, to create a flat ontology-like structure. Using this, images related to the same noun but for different adjectives, and vice versa, can be filtered. Each ANP comes with 21 sentiments with their probability (e.g., *joy* = 0.6, *ecstasy* = 0.8,) but all images with the same ANP share the same sentiment score. Sentiment scores per image are available through a second dataset [81], but they are on a single axis from positive to negative and only available for a small number of images. Therefore, in the following, per-image psycholinguistic labels are computed from the textual metadata. Each image also comes with textual metadata containing a title, a description text, and tags. This metadata is used in the following section to compute an individual psycholinguistic grounding for each image.

A.2.1.2 Per-image psycholinguistic scores

To create an embedding with a meaningful spatial distribution per image, individual scores for each image are needed. Here, a psycholinguistic grounding of the textual metadata for each image is calculated. Scott et al. [73] provide a psycholinguistics dataset with nine ratings each for 5,500 words. The nine ratings available are *arousal*, *dominance*, *valence*, *imageability*, *concreteness*, *familiarity*, *semantic size*, *age of acquisition*, and *gender association*. For each image, the title, description, and tags from the MVSO dataset are extracted. All these data are provided by the

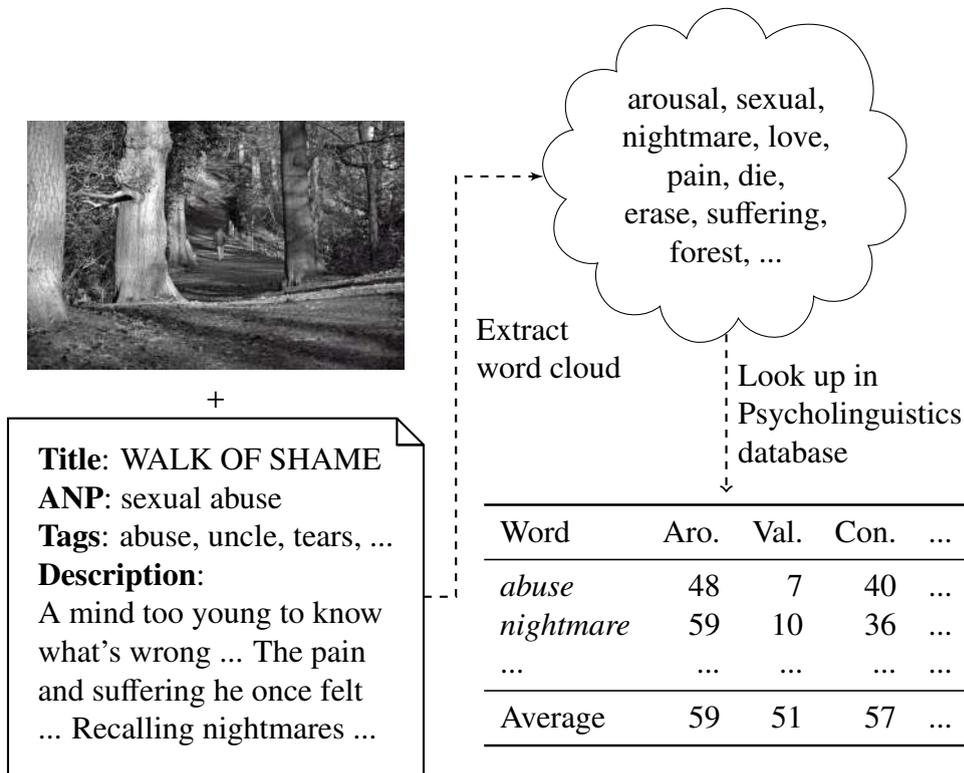


Figure A.5: Process of calculating per-image psycholinguistic scores.

image uploader, which makes them noisy. A word-cloud from all words used in the metadata is generated, stripping grammatical affixes through lemmatization. Furthermore, all words not contained in the psycholinguistics database are filtered out. Lastly, nine psycholinguistic ratings by averaging the corresponding scores for each word in the word-cloud are calculated. The process of calculating per-image psycholinguistic scores is shown in Fig. A.5¹. Filtering out images where the wording used in the meta-data was not available in the psycholinguistics dictionary, this results in approximately 400,000 images with nine individual psycholinguistic ratings each.

For each noun, adjective, and ANP, a spatial embedding is computed using UMAP [74]. Additionally, an embedding including all images, filtering for extreme cases with very high or very low scores for some psycholinguistic ratings is computed. As input, a 30-dimensional vector is used for each image, composed of the 21 sentiment

¹The example photo is courtesy of [despitestraightlines](#) [34].

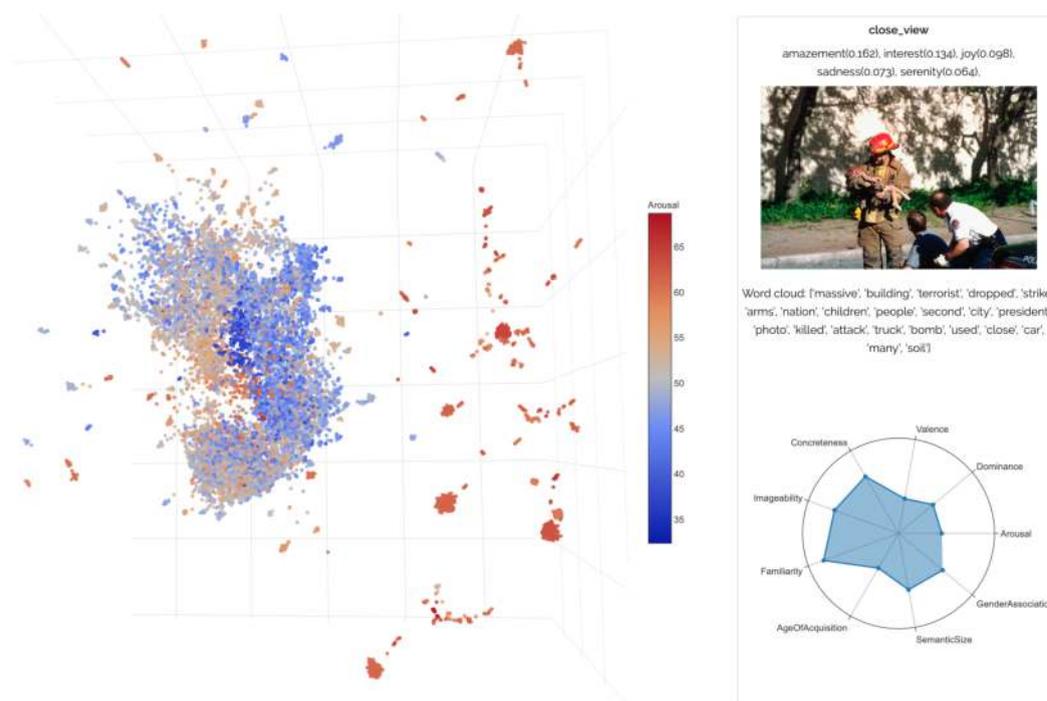


Figure A.6: Main user interface of the proposed visualization.

scores of its ANP as well as the nine psycholinguistic ratings calculated through the metadata.

A.2.2 Visualization

A dataset browser is built to visualize the relationship between human sentiment ratings of an image and the psycholinguistic characteristics of words used in the image metadata. Using this tool, we can browse the dataset, filter it for different adjectives or nouns, and see the scoring for different images. A three-dimensional view shows the sentiment-psycholinguistic spatial embedding of the selected dataset. Different color modes allow for analyzing the dataset regarding its ontology and human perception scores established in Section A.2.1. The full user interface of the proposed tool is shown in Figure A.6.

The sentiment-psycholinguistic space is shown with an interactive interface allowing for zooming and panning. Each data-point represents one image from the MVSO

dataset plotted on a three-dimensional embedding based on its individual psycholinguistic scores. We can switch between sampling a selection of images across the whole dataset, or showing all images of a selected noun, adjective, or ANP.

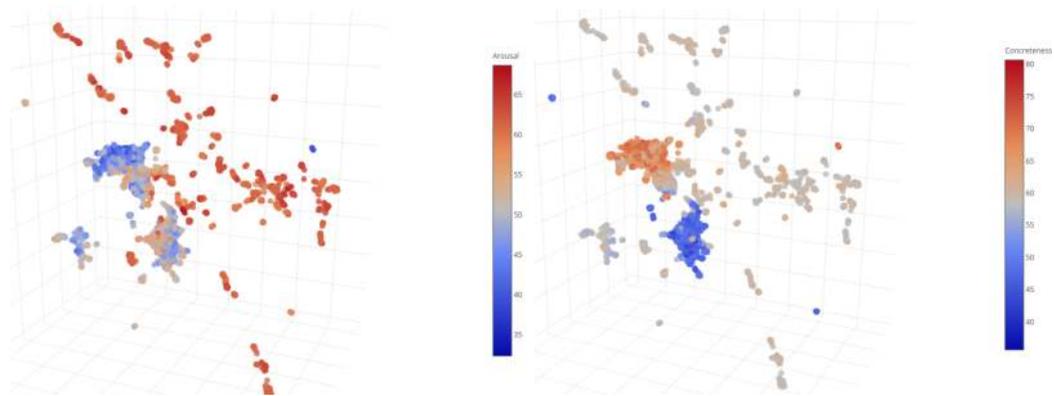
The color displayed in the spatial embedding can be selected to either show scores related to human perception as heatmap-based color gradings, or highlight the ontology-based class labels (e.g., *different adjectives* for a filtered *noun* dataset.) The different color modes are shown in Figure A.7.

When selecting a data sample in the spatial embedding, a detailed view opens on the right. Here, one can see the actual image behind the sample, as well as some of its metadata related to the sentiment score. A table shows the computed psycholinguistic values, as well as its highest and lowest significant words for each rating.

A.2.3 Summary

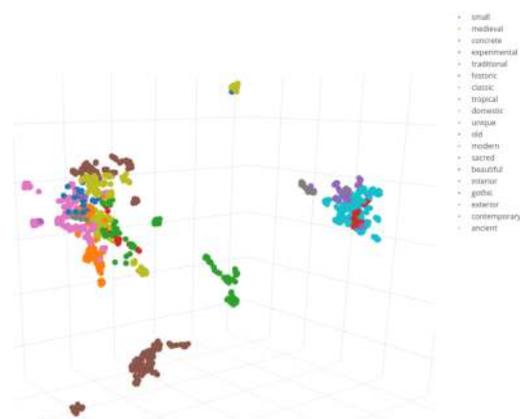
A tool to visualize sentiment image datasets regarding their psycholinguistic grounding has been introduced in this section. For each image, nine individual psycholinguistic scores are computed using textual metadata. A spatial embedding is computed to visualize their relationship of text and image. The interactive tool showcases the MVSO dataset, either wholly or by filtering it for nouns, adjectives, or ANPs. The spatial embedding gives further insights into how images for the same noun form different clusters regarding their human perception. Different color modes can be used to either highlight a single sentiment or psycholinguistic rating or visualize the ontology of the dataset.

As future directions, one could compare the visual characteristics of different clusters similar to the BoVW keypoint visualization used in Sec. A.1. Furthermore, the use of visual information to detect per-image sentiment scores could give additional insights into the perception of individual images. Lastly, as the MVSO dataset includes data in multiple languages, the visualization could be extended to work across multiple languages.



(a) Arousal score

(b) Concreteness score



(c) Ontology visualization

Figure A.7: Spatial embedding can be colored in different ways based on their calculated individual scores or dataset annotations.

Bibliography

- [1] Itseez. Open source computer vision library, 2015. URL: <https://opencv.org/>.
- [2] New Mexico State University. PSY301: Introduction to psycholinguistics, 2019. URL: <https://www.coursehero.com/sitemap/schools/1490-New-Mexico-State-University/courses/3064729-PSY301/>. Accessed on November 26, 2019.
- [3] S. Tanaka, A. Jatowt, M. P. Kato, and K. Tanaka. Estimating content concreteness for finding comprehensible documents. In *Proc. 6th ACM Int. Conf. on Web Search and Data Mining*, pages 475–484, February 2013. DOI: 10.1145/2433396.2433455.
- [4] J. Tang, X. Shu, G. Qi, Z. Li, M. Wang, S. Yan, and R. Jain. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(8):1662–1674, August 2017. DOI: 10.1109/TPAMI.2016.2608882.
- [5] J. Charbonnier and C. Wartena. Predicting word concreteness and imagery. In *Proc. 13th Int. Conf. on Computational Semantics*, pages 176–187, May 2019.

- [6] J. Hewitt, D. Ippolito, B. Callahan, R. Kriz, D. T. Wijaya, and C. Callison-Burch. Learning translations via images with a massively multilingual image dataset. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics, vol. 1*, pages 2566–2576, July 2018. DOI: [10.18653/v1/P18-1239](https://doi.org/10.18653/v1/P18-1239).
- [7] J. Hessel, D. Mimno, and L. Lee. Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proc. 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1*, pages 2194–2205, June 2018. DOI: [10.18653/v1/N18-1199](https://doi.org/10.18653/v1/N18-1199).
- [8] N. Ljubešić, D. Fišer, and A. Peti-Stantić. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proc. 3rd Workshop on Representation Learning for NLP*, pages 217–222, July 2018. DOI: [10.18653/v1/W18-3028](https://doi.org/10.18653/v1/W18-3028).
- [9] A. Vempala and D. Preoțiuc-Pietro. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840, July 2019. DOI: [10.18653/v1/P19-1272](https://doi.org/10.18653/v1/P19-1272).
- [10] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. From large scale image categorization to entry-level categories. In *Proc. 2013 IEEE Int. Conf. on Computer Vision*, pages 2768–2775, December 2013. DOI: [10.1109/ICCV.2013.344](https://doi.org/10.1109/ICCV.2013.344).
- [11] F. Smolik and A. Kriz. 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, October 2015. DOI: [10.1177/0142723715609228](https://doi.org/10.1177/0142723715609228).
- [12] S. Bai and S. An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, October 2018. DOI: [10.1016/j.neucom.2018.05.080](https://doi.org/10.1016/j.neucom.2018.05.080).

- [13] M. Vidanapathirana. YOLO3-4-Py, 2018. URL: <https://github.com/madhawav/YOLO3-4-Py>. Accessed on November 26, 2019.
- [14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, November 2011.
- [15] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *Computing Research Repository*, *arXiv:1707.02968*, August 2017.
- [16] D. Comaniciu and P. Meer. Mean Shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, August 2002. DOI: 10.1109/34.1000236.
- [17] Microsoft. Microsoft Azure Bing Search API, 2016. URL: <https://azure.microsoft.com/ja-jp/services/cognitive-services/search/>. Accessed on November 26, 2019.
- [18] H. Kawakubo, Y. Akima, and K. Yanai. Automatic construction of a folksonomy-based visual ontology. In *Proc. 2010 IEEE Int. Symposium on Multimedia*, pages 330–335, December 2010. DOI: 10.1109/ISM.2010.57.
- [19] D. Kiela, F. Hill, A. Korhonen, and S. Clark. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*, pages 835–841, June 2014. DOI: 10.3115/v1/P14-2135.
- [20] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev. To search or to label?: Predicting the performance of search-based automatic image classifiers. In *Proc. 8th ACM Int. Workshop on Multimedia Information Retrieval*, pages 249–258, October 2006. DOI: 10.1145/1178677.1178712.

- [21] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *Proc. 18th Int. Conf. on World Wide Web*, pages 341–350, April 2009. DOI: 10.1145/1526709.1526756.
- [22] N. Inoue and K. Shinoda. Adaptation of word vectors using tree structure for visual semantics. In *Proc. 24th ACM Multimedia Conf.*, pages 277–281, October 2016. DOI: 10.1145/2964284.2967226.
- [23] K. Nakamura and N. Babaguchi. Inter-concept distance measurement with adaptively weighted multiple visual features. In *Computer Vision —ACCV 2014 Workshops*, volume 9010 of *Lecture Notes in Computer Science*, pages 56–70. Springer, April 2015. DOI: 10.1007/978-3-319-16634-6_5.
- [24] F. Chollet et al. Keras. <https://github.com/fchollet/keras/>, 2015. Accessed on November 26, 2019.
- [25] J. D. J. Deng, W. D. W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. 2009 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 2–9, June 2009. DOI: 10.1109/CVPR.2009.5206848.
- [26] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. ECCV 2004 Workshop on Statistical Learning in Computer Vision*, pages 1–22, May 2004.
- [27] K. Yanai and K. Barnard. Image region entropy: A measure of “visualness” of Web images associated with one concept. In *Proc. 13th ACM Multimedia Conf.*, pages 419–422, November 2005. DOI: 10.1145/1101149.1101241.
- [28] A. Paivio, J. C. Yuille, and S. A. Madigan. Concreteness, imagery, and meaningfulness values for 925 nouns. *J. Exp. Psychol.*, 76(1):1–25, January 1968.
- [29] Y. Nagasawa, K. Nakamura, N. Nitta, and N. Babaguchi. Effect of junk images on inter-concept distance measurement: Positive or negative? In *Advances in*

- Multimedia Modeling: 23rd Int. Conf. on Multimedia Modeling Procs.*, volume 10133 of *Lecture Notes in Computer Science*, pages 173–184. Springer, December 2017. DOI: 10.1007/978-3-319-51814-5_15.
- [30] M. J. Cortese and A. Fugett. Imageability ratings for 3,000 monosyllabic words. *Behav. Res. Methods Instrum. Comput.*, 36(3):384–387, August 2004. DOI: 10.3758/BF03195585.
- [31] M. Zhang, R. Hwa, and A. Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proc. British Machine Vision Conf. 2018*, number 8, September 2018.
- [32] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. YFCC100M: The new data in multimedia research. *Comm. ACM*, 59(2):64–73, January 2016. DOI: 10.1145/2812802.
- [33] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of GIST descriptors for Web-scale image search. In *Proc. ACM Int. Conf. on Image and Video Retrieval 2009*, pages 19:1–19:8, 2009. DOI: 10.1145/1646396.1646421.
- [34] despitestraightlines. WALK OF SHAME (Alternate image), 2012. URL: <http://flickr.com/photos/despitestraightlines/6677983565/>. Accessed on November 26, 2019.
- [35] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. 10th Research on Computational Linguistics Int. Conf.*, pages 19–33, August 1997.
- [36] R. Cohendet, C.-H. Demarty, N. Q. K. Duong, M. Sjöberg, B. Ionescu, T.-T. Do, and F. Rennes. Mediaeval 2018: Predicting media memorability task. *Computing Research Repository*, arXiv:1807.01052, July 2018.
- [37] Google. Google Custom Search API, 2016. URL: <https://developers.google.com/custom-search/>. Accessed on November 26, 2019.

- [38] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovávr, J. Michelfeit, P. Rychlý, and V. Suchomel. The sketch engine: Ten years on. *Lexicography*, 1(1):7–36, July 2014. DOI: 10.1007/s40607-014-0009-9.
- [39] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, February 2003. DOI: 10.1162/153244303322533214.
- [40] M. Davies. The corpus of contemporary American English: 520 million words, 1990–present, 2008. URL: <http://corpus.byu.edu/coca/>. Accessed on November 26, 2019.
- [41] Y. Kohara and K. Yanai. Visual analysis of tag co-occurrence on nouns and adjectives. In *Advances in Multimedia Modeling: 19th Int. Conf. on Multimedia Modeling Procs.*, volume 7732 of *Lecture Notes in Computer Science*, pages 47–57. Springer, January 2013. DOI: 10.1007/978-3-642-35725-1-5.
- [42] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3270–3277, June 2014. DOI: 10.1109/CVPR.2014.412.
- [43] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008. DOI: 10.1016/j.cviu.2007.09.014.
- [44] E. Loper and S. Bird. NLTK: The Natural Language Toolkit. In *Proc. ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics Vol. 1*, pages 63–70, 2002. DOI: 10.3115/1118108.1118117.
- [45] G. A. Miller. WordNet: A lexical database for English. *Comm. ACM*, 38(11): 39–41, November 1995. DOI: 10.1145/219717.219748.
- [46] A. Budanitsky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proc. 2nd Meeting of*

- the North American Chapter of the Association for Computational Linguistics*, pages 29–34, June 2001.
- [47] L. L. Thurstone. The method of paired comparisons for social values. *J. Abnorm. Psychol.*, 21(4):384–400, 1927.
- [48] L. Maystre. Choix —Inference algorithms for models based on Luce’s choice axiom, 2017. URL: <https://github.com/lucasmaystre/choix/>. Accessed on November 26, 2019.
- [49] Merriam-Webster. Merriam-Webster Online Dictionary, 2017. URL: <http://www.merriam-webster.com/>. Accessed on November 26, 2019.
- [50] Oxford University Press. OED Online, 2017. URL: <https://en.oxforddictionaries.com/>. Accessed on November 26, 2019.
- [51] V. Coltheart, V. J. Laxon, and C. Keating. Effects of word imageability and age of acquisition on children’s reading. *Br. J. Psychol.*, 79(1):1–12, February 1988. DOI: 10.1111/j.2044-8295.1988.tb02270.x.
- [52] B. Giesbrecht, C. C. Camblin, and T. Y. Swaab. Separable effects of semantic priming and imageability on word processing in human cortex. *Cereb Cortex*, 14(5):521–529, May 2004. DOI: 10.1093/cercor/bhh014.
- [53] G. V. Jones. Deep dyslexia, imageability, and ease of predication. *Brain. Lang.*, 24(1):1–19, January 1985. DOI: 10.1016/0093-934X(85)90094-X.
- [54] P. J. Schwanenflugel. Why are abstract concepts hard to understand? In *The Psychology of Word Meanings*, pages 235–262. Psychology Press, New York, NY, USA, 2013.
- [55] W. Ma, R. M. Golinkoff, K. Hirsh-Pasek, C. McDonough, and T. Tardif. Imageability predicts the age of acquisition of verbs in Chinese children. *J. Child. Lang.*, 36:405–423, March 2009. DOI: 10.1017/S0305000908009008.

- [56] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.*, 6(1):3–5, February 2011. DOI: 10.1177/1745691610393980.
- [57] Y. Dodge. Spearman rank correlation coefficient. In *The Concise Encyclopedia of Statistics*, pages 502–505. Springer, New York, NY, 2008. DOI: 10.1007/978-0-387-32833-1_379.
- [58] J. Reilly and J. Kean. Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. *J. Cogn. Sci.*, 31(1):157–168, February 2010. DOI: 10.1080/03640210709336988.
- [59] A. Sianipar, P. van Groenestijn, and T. Dijkstra. Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Front. Psychol.*, 7:1907, December 2016. DOI: 10.3389/fpsyg.2016.01907.
- [60] L. T. S. Yee. Valence, arousal, familiarity, concreteness, and imageability ratings for 292 two-character Chinese nouns in Cantonese speakers in Hong Kong. *PloS one*, 12(3):e0174569, March 2017. DOI: 10.3389/fpsyg.2016.01907.
- [61] J. W. Pennebaker, M. E. Francis, and R. J. Booth. *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Erlbaum, Mahwah, NJ, USA, 2001.
- [62] A. Balahur, S. M. Mohammad, V. Hoste, and R. Klinger, editors. *Proc. 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium, 2018. Association for Computational Linguistics.
- [63] E. Fast, B. Chen, and M. S. Bernstein. Empath: Understanding topic signals in large-scale text. *Computing Research Repository, arXiv:1602.06979*, February 2016.
- [64] M. Coltheart. The MRC psycholinguistic database. *Q. J. Exp. Psychol. A.*, 33(4):497–505, January 1981. DOI: 10.1080/14640748108400805.

- [65] J. Redmon and A. Farhadi. YOLO9000: Better, faster, stronger. *Computing Research Repository*, arXiv:1612.08242, December 2016.
- [66] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001. DOI: 10.1023/A:1010933404324.
- [67] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. Ciftçi, H. Güleç, A. A. Salah, and M. Pantic, editors. *Proc. 2018 Audio/Visual Emotion Challenge and Workshop*, New York, NY, USA, 2018. ACM.
- [68] W. Samek, T. Wiegand, and K.-R. Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *Computing Research Repository*, arXiv:1708.08296, August 2017.
- [69] J. J. Li and A. Nenkova. Fast and accurate prediction of sentence specificity. In *Proc. 29th AAAI Conf. on Artificial Intelligence*, pages 2281–2287, January 2015.
- [70] C. Hentschel and H. Sack. What image classifiers really see —Visualizing bag-of-visual words models. In *Advances in Multimedia Modeling: 21st Int. Conf. on Multimedia Modeling Procs.*, volume 8935 of *Lecture Notes in Computer Science*, pages 95–104. Springer, January 2015. DOI: 10.1007/978-3-319-14445-0_9.
- [71] A. Holzinger, B. Malle, P. Kieseberg, P. M. Roth, H. Müller, R. Reihs, and K. Zatloukal. Towards the augmented pathologist: Challenges of explainable-ai in digital pathology. *Computing Research Repository*, arXiv:1712.06657, 2017.
- [72] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell. What do we need to build explainable AI systems for the medical domain? *Computing Research Repository*, arXiv:1712.09923, December 2017.

- [73] G. G. Scott, A. Keitel, M. Becirspahic, B. Yao, and S. C. Sereno. The Glasgow norms: Ratings of 5,500 words on nine scales. *Behav. Res. Meth.*, 51(3):1258–1270, June 2019. DOI: 10.3758/s13428-018-1099-3.
- [74] L. McInnes and J. Healy. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *Computing Research Repository*, *arXiv:1802.03426*, February 2018.
- [75] S. Jindal and S. Singh. Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning. In *Proc. 2015 IEEE Int. Conf. on Image Processing*, pages 447–451, December 2015. DOI: 10.1109/INFOP.2015.7489424.
- [76] E. Kim and R. Klinger. A survey on sentiment and emotion analysis for computational literary studies. *Computing Research Repository*, *arXiv:1808.03137*, August 2018.
- [77] Bokeh Development Team. *Bokeh: Python library for interactive visualization*, 2014. URL: <http://www.bokeh.pydata.org/>. Accessed on November 26, 2019.
- [78] J. T. E. Richardson. Imageability and concreteness. *Bull. Psychon. Soc.*, 7(5): 429–431, May 1976. DOI: 10.3758/BF03337237.
- [79] H. Yue, W. Chen, X. Wu, and J. Wang. Visualizing bag-of-words for high-resolution remote sensing image classification. *J. Appl. Remote Sens.*, 10(1), March 2016. DOI: 10.1117/1.JRS.10.015022.
- [80] M. Wilson. MRC psycholinguistic database: Machine-usable dictionary, version 2.00, January 1988.
- [81] V. D., H. Liu, and S.-F. Chang. Columbia MVSO image sentiment dataset. *Computing Research Repository*, *arXiv:1611.04455*, *arXiv:1611.04455*, November 2016.

- [82] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. *Computing Research Repository*, arXiv:1508.03868, August 2015.
- [83] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern. Anal. Mach. Intell.*, 22(12):1349–1380, December 2000. DOI: 10.1109/34.895972.
- [84] M. C. Tacca. Commonalities between Perception and Cognition. *Front. Psychol.*, 2:358, November 2011. DOI: 10.3389/fpsyg.2011.00358.
- [85] S. Cheng, W. Chen, and H. Sundaram. Semantic visual templates: Linking visual features to semantics. In *Proc. 1998 Int. Conf. on Image Processing*, volume 3, pages 531–535, October 1998. DOI: 10.1109/ICIP.1998.727321.
- [86] Y. Wang, N. Stash, L. Aroyo, L. Hollink, and G. Schreiber. Semantic relations in content-based recommender systems. In *Proc. 5th Int. Conf. on Knowledge Capture*, pages 209–210, September 2009. DOI: 10.1145/1597735.1597786.
- [87] R. Zhao and W. I. Grosky. Narrowing the semantic gap —Improved text-based Web document retrieval using visual features. *IEEE Trans. Multimed.*, 4(2):189–200, June 2002. DOI: 10.1109/TMM.2002.1017733.
- [88] R. Zhao and W. I. Grosky. Bridging the semantic gap in image retrieval. In *Distributed Multimedia Databases: Techniques and Applications*, pages 14–36, Hershey, USA, July 2002. Idea Group Publishing.
- [89] F. Nack, C. Dorai, and S. Venkatesh. Computational media aesthetics: Finding meaning beautiful. *IEEE MultiMed.*, 8(4):10–12, October 2001. DOI: 10.1109/93.959093.
- [90] C. Dorai and S. Venkatesh. Bridging the semantic gap with computational media aesthetics. *IEEE MultiMed.*, 10(2):15–17, April 2003. DOI: 10.1109/MMUL.2003.1195157.

- [91] Xiangbo Shu, Guo-Jun Qi, Jinhui Tang, and Jingdong Wang. Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation. In *Proc. 23rd ACM Multimedia Conf.*, pages 35–44, October 2015. DOI: [10.1145/2733373.2806216](https://doi.org/10.1145/2733373.2806216).
- [92] J. Tang, X. Shu, Z. Li, G.-J. Qi, and J. Wang. Generalized deep transfer networks for knowledge propagation in heterogeneous domains. *ACM Trans. Multimed. Comput. Commun. Appl.*, 12(68):68:1–68:22, November 2016. DOI: [10.1145/2998574](https://doi.org/10.1145/2998574).
- [93] J. Tang, X. Shu, Z. Li, Y. Jiang, and Q. Tian. Social anchor-unit graph regularized tensor completion for large-scale image retagging. *IEEE Trans. Pattern. Anal. Mach. Intell.*, 41(8):2027–2034, August 2019. DOI: [10.1109/TPAMI.2019.2906603](https://doi.org/10.1109/TPAMI.2019.2906603).
- [94] C. Montemayor and H. H. Haladjian. Perception and cognition are largely independent, but still affect each other in systematic ways: Arguments from evolution and the consciousness-attention dissociation. *Front. Psychol.*, 8:40, January 2017. DOI: [10.3389/fpsyg.2017.00040](https://doi.org/10.3389/fpsyg.2017.00040).
- [95] A. Cahen and M. C. Tacca. Linking perception and cognition. *Front. Psychol.*, 4:144, March 2013. DOI: [10.3389/fpsyg.2013.00144](https://doi.org/10.3389/fpsyg.2013.00144).
- [96] N. Dijkstra, S. E. Bosch, and M. A. J. van Gerven. Shared Neural Mechanisms of Visual Perception and Imagery. *Trends Cogn. Sci. (Regul. Ed.)*, 23(5):423–434, May 2019.
- [97] S. Dellantonio, R. Job, and C. Mulatti. Imageability: Now you see it again (albeit in a different form). *Front. Psychol.*, 5:279, April 2014. DOI: [10.3389/fpsyg.2014.00279](https://doi.org/10.3389/fpsyg.2014.00279).
- [98] J. Pearson, T. Naselaris, E. A. Holmes, and S. M. Kosslyn. Mental Imagery: Functional Mechanisms and Clinical Applications. *Trends Cogn. Sci. (Regul. Ed.)*, 19(10):590–602, October 2015.

-
- [99] J. T. E. Richardson. *Mental Imagery and Human Memory*. Macmillan, London, 1980.
- [100] C. Otto, S. Holzki, and R. Ewerth. “Is this an example image?” —Predicting the relative abstractness level of image and text. *Computing Research Repository*, arXiv:1901.07878, January 2019.
- [101] C. Otto, M. Springstein, A. Anand, and R. Ewerth. Understanding, categorizing and predicting semantic image-text relations. In *Proc. 2019 Int. Conf. on Multimedia Retrieval*, pages 168–176, June 2019. DOI: 10.1145/3323873.3325049.
- [102] M. G. Constantin, M. Redi, G. Zen, and B. Ionescu. Computational understanding of visual interestingness beyond semantics: Literature survey and analysis of covariates. *ACM Comput. Surv.*, 52(2):25:1–25:37, March 2019. DOI: 10.1145/3301299.

Publication list

Journal

- [1] M. A. Kastner, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, and H. Murase. Estimating the visual variety of concepts by referring to Web popularity. *Multimed. Tools Appl.*, 78(7):9463–9488, April 2019. DOI: 10.1007/s11042-018-6528-x.
- [2] M. A. Kastner, I. Ide, F. Nack, Y. Kawanishi, T. Hirayama, D. Deguchi, and H. Murase. Estimating the imageability of words by mining visual characteristics from crawled image data. *Multimed. Tools Appl.*, 33 pages, January 2020. DOI: 10.1007/s11042-019-08571-4.

International conference

- [1] M. A. Kastner, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, and H. Murase. Browsing visual sentiment datasets using psycholinguistic groundings. In *Advances in Multimedia Modeling: 26th Int. Conf. on Multimedia Modeling Procs.*, volume 11962 of *Lecture Notes in Computer Science*, pages 697–702. Springer, January 2020. DOI: 10.1007/978-3-030-37734-2_56.

- [2] M. A. Kastner. On quantizing the mental image of concepts for visual semantic analyses. In *Proc. 27th ACM Int. Conf. on Multimedia*, pages 1660–1664, October 2019. DOI: [10.1145/3343031.3352589](https://doi.org/10.1145/3343031.3352589).

Domestic conference

- [1] M. A. Kastner, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, and H. Murase. A preliminary study on estimating word imageability labels using Web image data mining. In *Proc. 25th Annual Meeting of The Association for Natural Language Processing, Japan*, pages 747–750, March 2019.
- [2] M. A. Kastner, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, and H. Murase. On understanding visual relationships of concepts by visualizing bag-of-visual-words models. In *Proc. 21st Meeting on Image Recognition and Understanding, Japan*, number PS3-34, August 2018.
- [3] M. A. Kastner, Y. Tsuchiya, K. Osumi, R. Akiyama, L. Kawai, and H. Ikeda. A Survey on Psychology —Connecting Perception, Language, and Memory studies with Computer Vision. In *Proc. 21st Meeting on Image Recognition and Understanding, Japan*, number WT-7, August 2018.
- [4] M. A. Kastner, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, and H. Murase. On visualizing psycholinguistic groundings for sentiment image datasets. In *Proc. 22nd Meeting on Image Recognition and Understanding, Japan*, number DS-3, August 2019.
- [5] 梅村和紀, カストナーマークアウレル, 井手一郎, 川西康友, 平山高嗣, 道満恵介, 出口大輔, 村瀬洋. 画像キャプションの質的評価に向けた文の心像性推定手法の検討. In *Proc. 25th Annual Meeting of The Association for Natural Language Processing, Japan*, pages 755–758, March 2019.

- [6] カスタナーマークアウレル, 井手一郎, 川西康友, 平山高嗣, 出口大輔, 村瀬洋. Web画像の分布に基づく単語概念の視覚的な多様性の推定. 情報処理学会研究報告, number 4, March 2018.