Estimating the imageability of words by mining visual characteristics from crawled image data

Marc A. Kastner, Ichiro Ide, Frank Nack, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi & Hiroshi Murase

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501
Volume 79
Combined 25-26

Multimed Tools Appl (2020)
79:18167-18199
DOI 10.1007/s11042-019-08571-4

Springer

Multimedia Tools and Applications (2020) 79:18167–18199
https://doi.org/10.1007/s11042-019-08571-4

Estimating the imageability of words by mining visual characteristics from crawled image data

Marc A. Kastner[1] · Ichiro Ide[1] · Frank Nack[2] · Yasutomo Kawanishi[1] · Takatsugu Hirayama[3] · Daisuke Deguchi[4] · Hiroshi Murase[1]

Received: 22 February 2019 / Revised: 3 October 2019 / Accepted: 6 December 2019 /
Published online: 29 February 2020

Abstract

Natural Language Processing and multi-modal analyses are key elements in many applications. However, the semantic gap is an everlasting problem, leading to unnatural results disconnected from the user's perception. To understand semantics in multimedia applications, human perception needs to be taken into consideration. Imageability is an approach originating from Pyscholinguistics to quantize the human perception of words. Research shows a relationship between language usage and the imageability of words, making it useful for multimodal applications. However, the creation of imageability datasets is often manual and labor-intensive. In this paper, we propose a method using image data mining of a variety of visual features to estimate the imageability of words. The main assumption is a relationship between the imageability of concepts, human perception, and the contents of Web-crawled images. Using a set of low- and high-level visual features from Web-crawled images, a model is trained to predict imageability. The evaluations show that the imageability can be predicted with both a sufficiently low error, and a high correlation to the ground-truth annotations. The proposed method can be used to increase the corpus of imageability dictionaries.

Keywords Image classification · Deep transfer network · Heterogeneous-domain knowledge propagation · Cross-domain label transfer

1 Introduction

Whether it is multimedia retrieval applications, consumer applications, recommendation engines, or Big Data analyses in Web and Social Media in general, the use of multiple

Parts of this research were supported by JSPS KAKENHI 16H02846, and a joint research project with NII, Japan.

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

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

modalities became ubiquitous for multimedia applications. However, the so-called *semantic gap* is an everlasting problem for various multimodal applications. As semantics are often hard to transfer between modalities, application results can often be perceived disconnected from the user, a human. Whether it is processing text or images, human perception and related semantics are often ignored. As many applications deal with language, it seems reasonable to include a metric of human perception into the processing of language.

Imageability is a concept originating from Psycholinguistics. 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 high-imageability word 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 low-imageability word 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 [8, 30], language understanding [39], and the use of grammar [42]. The concept of imageability, along with example images for differently imageable words, is visualized in Fig. 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 commonly used as a complementary feature for sentiment research [2, 37], but found its way into recent multimodal research using text and image [52]. 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.

leaf (imageability 6.08) *early* (imageability 3.90)

Fig. 1 Imageability of words. High-imageability words like *leaf* are concrete terms describing actual objects. Thus, when comparing images related to the same word, the images are visually similar. In contrast, low-imageability words like *early* are often concepts or abstract ideas. Images related to these words share fewer visual similarities. Imageability is commonly described as a seven-level Lickert scale ranging from very unimageable to very imageable

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, we propose a method using image-based data-mining to estimate the imageability of words. The core assumption is that imageability is a quantization of mental image of a certain word, describing how we perceive it. Thus, we further assume that there is an intrinsic relationship between the imageability of words, how we perceive the world around us, and how we capture this in images we upload to Social Media platforms.

Therefore, in our method, we first crawl large image sets for words for which we have ground-truth scores for imageability. 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, our 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, similarity matrices to describe the structural resemblance of all images in the same image sets, are calculated. Last, a model is trained to regress the imageability for unknown words. The model is evaluated using a series of testing data sets. In the experiments, we first evaluate the general performance of the proposed method in comparison to our previous work. Then, the feature selection gets a closer inspection, to investigate which features can excel for which type of word. Finally, we discuss some implications following the results of each experiment.

The paper is structured as follows. In Section 2, previous research on imageability, its applications, and our previous research on this topic is discussed. The core assumption of how image data crawled from the Web correlates to the human perception of imageability is discussed in Section 3, together with our proposed method and the used mixture of low-level and high-level visual image features. For the evaluation, we prepared a word corpus with imageability annotations, and a large set of images for up to 1,000 words, discussed in Section 4. Section 5 analyzes the proposed method in four experiments, looking at the choice of image features, the choice of regressor, dataset size, and how the choice of visual feature affects the performance for lowly or highly imageable words. Section 6 discusses the found results, as well as some implications for future applications, before concluding the paper in Section 7.

2 Related work

The concept of imageability and human perception in language understanding goes back to the 1960s, starting in the field of Psychology. From there, the concept naturally found its way into Psycholinguistics, Multimedia research, and Computer Science.

In the following, we will first look into research of imageability itself, starting with Psycholinguistics. Next, an overview on recent research applying psycholinguistic features in Multimedia (and Multimodal) research is outlined. In the end, other applications, where imageability research could have an impact, are discussed.

[1] https://www.mturk.com/

Psycholinguistics In 1968, Paivio et al. [32] first proposed the concepts of *imageability*, *concreteness*, and *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. The imageability of verbs has implications on grammar usage for different contexts [42], which could provide helpful knowledge to create more natural language depending on context. There is also a relationship on imageability of words to age of acquisition and reading comprehension, especially relevant for children [8, 30]. Due to this, there are further implications for research related to dyslexia [23]. There is a relationship of text difficulty and concreteness, when it comes to abstract words, as it represents the fundamental semantic distinction between them [39]. In Neuropsychology, there is research on the neurological process of word understanding in relation to their imageability [16]. There are imageability dictionaries for English [10, 36] as well as other languages [41, 51]. However, the dictionary creation process is labor-intensive, as the annotations are commonly obtained through crowd-sourcing or user studies involving test subjects.

Visual concept analysis In Multimedia research, the analysis of visual concepts have been ground for multiple works. Prominently, this research involves estimating or quantifying relationship of different concepts. Therefore, it derives hierarchical structures or ontologies [21, 25] of concepts from their visual relationships. Other work by Yanai and Barnard [50] analyzed image region entropy to identify *visualness* of adjectives, later continued by Kohara and Yanai [26] to analyze adjective-noun pairs. Divvala et al. [13] proposed a method to analyze visual features to create visual knowledge databases with unsupervised crawling. Tang et al. [46] look at social-aware tagging by including user-information into the training to remove noisy and unimportant tags.

Imageability estimation In the field of Natural Language Processing, researchers have been working towards the estimation of imageability or concreteness using text data mining techniques. Ljubesic et al. [28] create 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. Hessel et al. [17] use the multimodal abstractness of concepts to learn better image/text correspondences. They conclude an improved retrieval performance through the introduction of concreteness and imageability in word embeddings of multimodal data sets. In a similar sense, Hewitt et al. [18] use the concreteness of concepts across multilingual image datasets to improve the results of translations.

For our work, we were interested in how much knowledge of imageability can be gained from just analyzing the visual characteristics of image datasets. In our previous work [24], we proposed the idea to estimate the *visual variety* of terms as a measurement of abstractness. The idea is to quantize the mental image of different concepts, based on the variety in their visual characteristics. The proposed method is a data-driven approach which first creates *ideal* image sets using recomposition of existing datasets. Then, a naïve clustering-based approach is applied on the visual features to determine the variety. The evaluation covered a small number of 25 terms related to vehicles, which led to promising results for estimating variety gaps within the same domain. However, as the method uses a single visual feature, the encoded visual vectors are not exhaustive enough to compare data across differ-

ent domains. Additionally, experiments on a larger scale turned out to be unfeasible due to limitations in the data acquisition process. To the best of our knowledge, there has been no similar research which employs visual concept analyses for imageability estimation. While our previous work looked at visual variety of related concepts, a relationship between this and the concept of imageability is undeniable. Thus, in this paper, we will apply similar ideas to the concept of imageability. As such, we propose a more sophisticated data-mining approach using a variety of visual features to encode the visual characteristics of each word from various angles, and then train a model to estimate imageability scores for words based on ground-truth data.

As the use of imageability for multimedia applications has been evaluated before [17, 18, 52], we foremost focus on the actual estimation of imageability labels for extending existing datasets, evaluating our results against ground-truth data from Psycholinguistics. Furthermore, we adapt Ljubesic et al.'s method [28] as a comparison method to show how the results of mining textual data differs to visual data.

Multimodal applications Some works in the previous paragraph already introduced use-cases of concreteness for retrieval and translation applications.

Tanaka et al. [44] use content concreteness of documents to find comprehensible documents, finding a positive correlation between concreteness and content comprehensibility. Furthermore, there is also research in using deep networks to model cross-domain information between text and images [40, 45, 47].

Other opportunities of imageability are not yet heavily researched for multimedia purposes. However, there have been some applications using Psycholinguistic metrics as complementary features in recent work. Zhang et al. [52] analyzed the implicit relationship of image and text for posters and advertisements. They look 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 tries 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 intrinstic semantic relationships.

Another typical use-case for Psycholinguistic features is sentiment and emotion analysis. 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 [34] and Empath [15], 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 of human perception. As such, it became the topic of regular workshops for both Multimedia [37] and Natural Language Processing conferences [2].

Other applications Other than multimodal applications and sentiment, Psycholinguistic features can be used in a number of other fields, too. Computational linguistics, or Natural Language Processing on its own, can profit from such metrics as a complement to sentiment embeddings. Li and Nenkova [27] use *imageability*, *concreteness*, and *meaningfulness* to

predict sentence specificity. The proposed method can be used to estimate text difficulty or create simplified versions of text.

There is also the recently established new field called Explainable AI [38] (XAI). In XAI, the goal is to get an better understanding on 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 mis-classifications. As many multimedia applications use neural networks for classification of language, be it personal assistants or translation tools, an additional insight on human perception can help to explain mis-classifications or unnatural results. There have been analyses related to Explainable AI for the fields of aviation and medicine, where a faulty classification could be potentially fatal [19, 20]. 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 blackboxed models involving language and vision.

In summary, *imageability* and similar numerical metrics for quantizing human perception have been part of Psychology research since the 1960s. In recent development of multimedia applications, such metrics are becoming more and more ubiquitous for multimodal applications as supplementary features for capturing the human perception through multiple modalities. An example of this is the detection of emotions and sentiments, often in the context of Web or Social Media data. With the rise of neural networks as the most common prediction model, Explainable AI is becoming a newly established field which yearns for additional insight on the hidden semantic relationships.

3 Imageability estimation

In this paper, we propose a method to estimate the imageability of words using visual feature mining on Web-crawled images. In the core assumption, we consider that there is an intrinsic relationship between imageability scores and the perceived world around us. This relationship is 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 capture 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 represented in different visual 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 thus the imageability of it.

In our previous research [24], we looked at the *visual variety* of related words. The proposed method recomposited a custom dataset to contain images in the same ratio of subconcepts as they would exist in real life. A dataset of *vehicles* would contain as many *cars*, compared to *airplanes* and *ships*, as these ratios would be in the perception of a common human being. Then, a simple data-mining approach using Mean Shift clustering [9] on local feature descriptors is applied on the created datasets to estimate the *visual variety* of the said dataset. The research was evaluated with 25 words related to vehicles and a small dataset of 2,400 images for each word. While the results looked promising within the same domain, the proposed data-mining method is not exhaustive enough to compare words across domains. In this paper, we shift the focus 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, we propose a sophisticated method using a cross comparison of similarity between all images in the dataset of a word. 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.

3.1 Approach

For now, we 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 values), and rather abstract words or concepts (usually low values).

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. High-imageability words are expected to have a high similarity across all their images. In contrast, Low-imageability words are expected to have a significantly lower similarity across their images.

Using a variety of visual features (Discussed in Section 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 values.

The step-by-step algorithm is shown in Fig. 2 and further described in Algorithm 1.

Algorithm 1 Pseudo-code for the proposed method.

input : Word
output: Imageability label

(1) data preparation;
$images \longleftarrow$ image dataset;
$words \longleftarrow$ psycholinguistics dataset;
$features \longleftarrow$ list of visual features;
for $image \in images$ **do**
  $image\_text \longleftarrow$ read textual metadata of $image$;
  if $image\_text \cap words \neq \emptyset$ **then**
    for $word \in image\_text \cap words$ **do**
      $images_{word} \longleftarrow image$;
    end
  end
end
(2) feature extraction;
for $word \in words$ **do**
  for $image \in images_{word}$ **do**
    for $feature \in features$ **do**
      $images_{word,feature,image} \longleftarrow$ extract visual features;
    end
  end
  for $feature \in features$ **do**
    $similarity\_matrix_{word,feature} \longleftarrow$
    cross comparison similarity for all in $images_{input,feature}$;
  end
end
(3a) training;
for $word \in words$ **do**
  $X_{word} \longleftarrow \|_{i=1}^{m}$ Eigenvalues of $similarity\_matrix_{word,i}$(for m features);
  $Y_{word} \longleftarrow words_{word}$;
end
train regression model Y on X;
(3b) prediction;
$X \longleftarrow \|_{i=1}^{m}$ Eigenvalues of $similarity\_matrix_{input,i}$(for m features);
predict Y from X;
$output \longleftarrow Y$;

3.2 Feature selection

To sufficiently encode the visual characteristics of each image set, 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

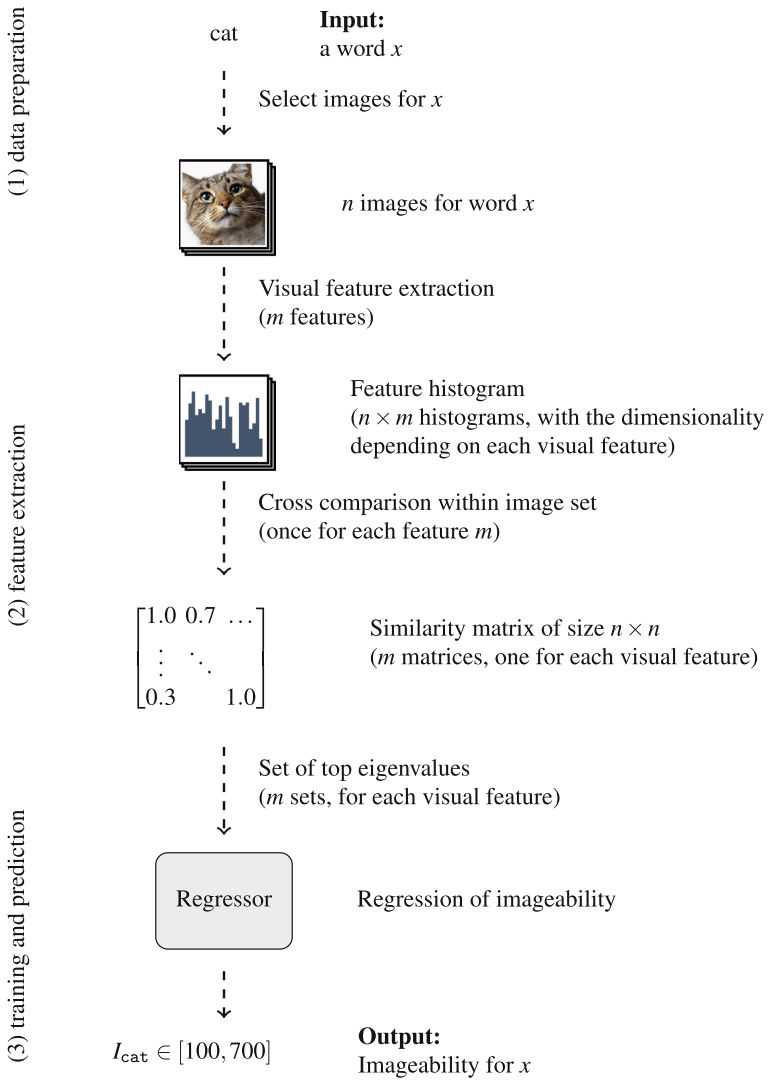

Fig. 2 Flowchart of the imageability estimation process. For each word, its corresponding image set gets analyzed based on a set of low- and high-level visual features. A cross comparison of all images is performed to create a separate similarity matrix for each visual feature. The eigenvalues of the similarity matrices are used to regress an imageability score. Vertical labels on the left refer to the corresponding section in Algorithm 1

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, we are interested in the actual image contents and compositions. The features are used to complement the visual feature representations in *what* and *how* things are displayed in each image, while putting the actual technical details (e.g. low-level details) to the side.

3.2.1 Low-level features

Low-level features look at 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 this paper, 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 [11] 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 5.1.

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 itself. 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 this paper, we investigate three characteristics of high-level representations:

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, while having different to colors or patterns, two images are perceived rather similar to humans if one contains a black small nude cat and another contains a white large fluffy cat.

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 3.1 above. Actual implementation details in which models are used for the evaluations are given in Section 5.1.

4 Dataset

In our research, we employ two types of datasets. First, a dictionary with English (language) words and imageability annotations, which provides the ground truth for both the training process and the evaluation. Second, a large number of images for each word, which will be used for visual feature extraction.

4.1 Imageability dictionary

There are a number of imageability or concreteness dictionaries in different languages, including English [10, 36], Indonesian [41], and Cantonese [51]. 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's 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 pair 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, we will look at the English language. Concretely, we use the datasets by Reilly et al. [36] and Cortese et al. [10] as a baseline. These datasets provide the results as a Likert scale averaged over all test subjects, in the range of [100, 700]. 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 [7] has not been used directly, despite it being larger than the previously cited sources. However, the first dataset used [36] is a modified version of the MRC data. It focuses on the high- and low-end of

the spectrum, removing large parts of mid-Imageability terms from the dictionary. In that process, they also filtered out obscure and uncommon terms, making for a cleaned-up fork of the MRC data.

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, we take the average of both dictionaries.

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, we normalize the interval of [100, 700] to [0, 100].

4.2 Image sets

In previous research [24], we looked at the task of measuring the visual variety of concepts using a dataset-driven approach. While the actual relationship of visual variety to imageability measurements remains to be verified, they are presumed to be similar. The approach looked at how the mental image of *vehicle* is related to its subordinate concepts like *car*, *boat*, or *plane*. The core assumption was, that the ratio of such sub-concepts relates to how humans create a mental image of the parent concept, as a sub-concept daily seen in daily life (*car*) may have a stronger influence than a concept rarely seen (*jet*). Thus, for each concept, a custom dataset was created using the WordNet [31] hierarchy of its hypernyms and a popularity measurement to determine the importance of sub-concept images in its parent datasets. The resulting dataset was considered to be *ideal*, meaning that the dataset composition resembles the frequency of subordinate concepts in real life, which was assumed to directly relate to the visual variety of the parent concept. While the approach led to promising results, it came with several downsides:

- The number of images available for very obscure sub-concepts could heavily bottleneck the re-composition 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.
- As it was tied to WordNet [31] and ImageNet [12], it would not work for words which were not available in both. It also fully relied on a hierarchy of hypernyms and hyponyms not available for all terms. Furthermore, ImageNet is rather limited in terms of both term availability and image availability, and only provides image data for nouns.
- 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.
- The process was mainly about the re-composition of parent concepts, so the most-bottom subordinate concepts would not benefit from the majority of the proposed contributions.

Due to these limitations, prior evaluations were only performed on a rather limited dataset of 25 terms related to vehicles, and about 2,400 images each.

Therefore, for this research, we propose a simplified method crawling image data from Social Media platforms for each word directly. The whole process of dataset acquisition is shown in Fig. 3. The crowd-sourced nature of our noisy Web-based origin dataset ensures a composition which comes close to how a human perceives the concept. The simplicity of direct crawling, on the other hand, ensures that we can retrieve a larger number of images for a much larger number of words. Therefore, we can evaluate the stability of our proposed algorithm from Section 3 with a large number of words. As the proposed dataset creation

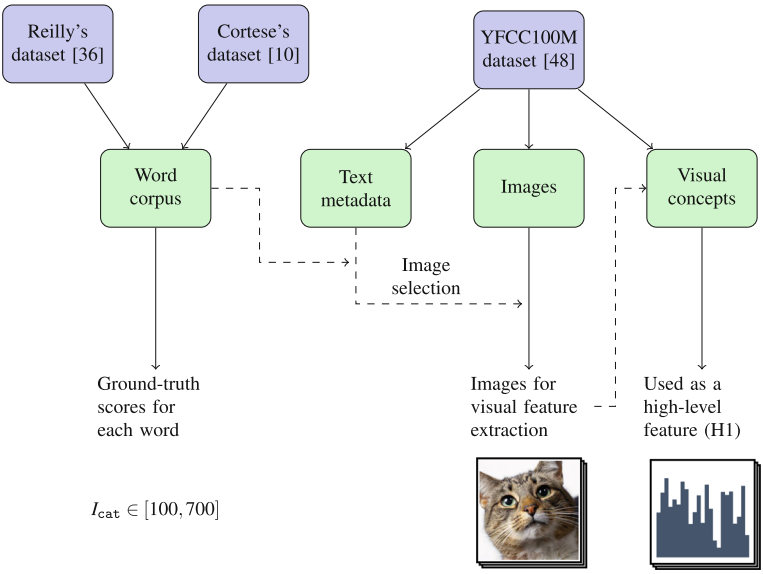

Fig. 3 Flowchart of dataset acquisition. The blue boxes show external origin datasets, while the green boxes indicate data crawled from these

method does not rely on WordNet, it implicitly groups ambiguous terms, and it 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 4.1 as a basis, a large number of images for each word with imageability annotation is crawled. As a source for the images, we use the YFCC100M [48] dataset, which is crowd-sourced based on the US photography social media platform Flickr[2]. YFCC100M 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 classification. This can be used as a high-level feature on its own and will be discussed later. For our research, we use the images themselves for visual feature data mining. Furthermore, the text-based annotations are used to identify a 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, we crawled the YFCC100M dataset, looking for images where entries from our imageability dictionaries appear in the text annotations. In case of multiple related words, the image is flagged to be part of the image set 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

[2] https://www.flickr.com/

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 our 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 in the next section.

5 Evaluation

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

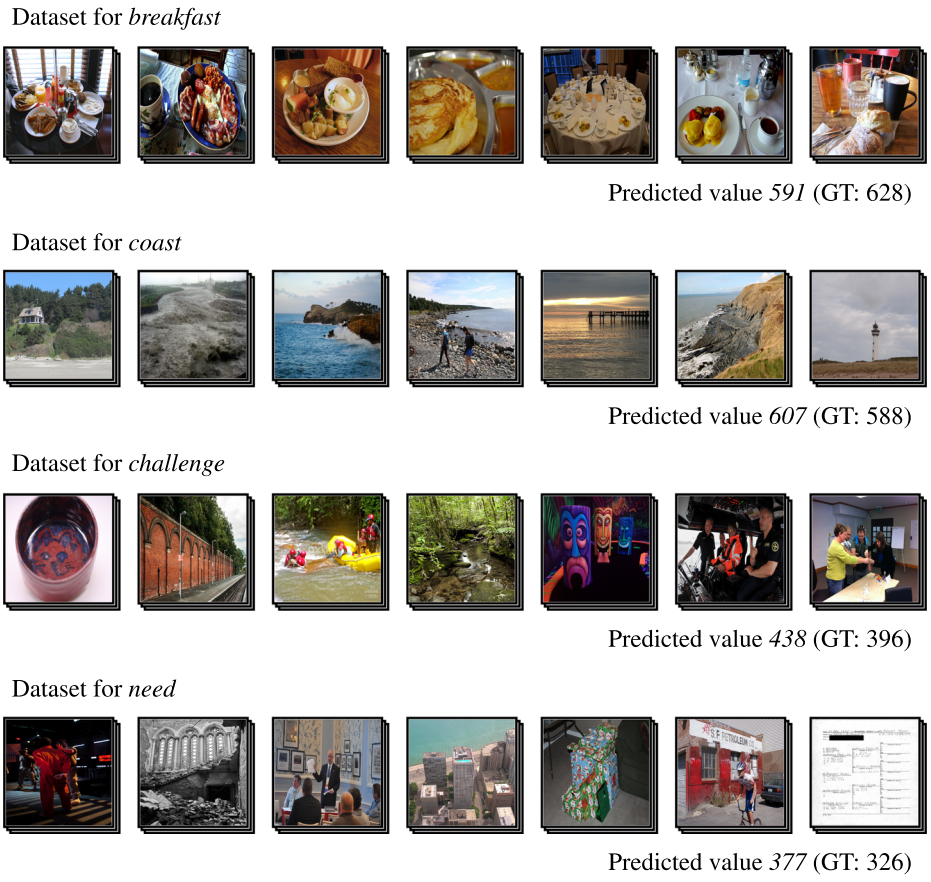

Fig. 4 Example of image datasets and their predicted imageability. For high-imageability words like *breakfast*, the resulting dataset is rather homonomous, having many similar scenery or objects in each image. In contrast, low-imageability words like *need* result in rather noisy datasets, often not clear why images belong to the dataset due to the vagueness of these abstract concepts. This noise is expected and used by the proposed method to predict a fitting imageability score. The predicted results are in the range [100, 700]

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

5.1 Environment

5.1.1 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 [3] to generate a Bag-of-Words model [11] 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 [14] commonly used for scene analysis. Based on this global gradient encoding, we generate a 960-dimensional histogram for each image.

Second, 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. We use the YFCC100M-based autotaggings provided by the dataset (as shown in Fig. 3). 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) This *Image content* feature encodes objects in each image. We use the pretrained model YOLO9000 [35] 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 [31] is used to group classes along their hypernyms. The dimensionality is 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, we

generate a histogram describing the number of objects within each grid cell. Here, n is set to 10, resulting in a 100-dimensional histogram.

Each feature is used to calculate a similarity matrix as outline in Section 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, we use the top 30 eigenvalues of each visual feature 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 [22] is used. For YOLO9000, the Python implementation YOLO3-4-Py [49] is used. For histogram comparisons, the normalized cross-correlation metric is used.

5.1.2 Dataset

Following the process discussed in Section 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 obtained 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 we 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. We found that this gives us a good balance of a sufficient number of images for data-mining, while having a sufficient 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-speech [3]. 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 data set is also shown in Fig. 5 in Section 5.2, together with results for the proposed and comparative methods.

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

Example images from the created image dataset are shown in Fig. 4.

5.1.3 Regression model

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

The evaluations, if not indicated otherwise, use Random Forest [4] 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 [33], while the latter is implemented in Keras 2.0.6 [6].

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.

[3] Parts-of-speech are obtained using NLTK [29] and may thus have slight error due to ambiguities.

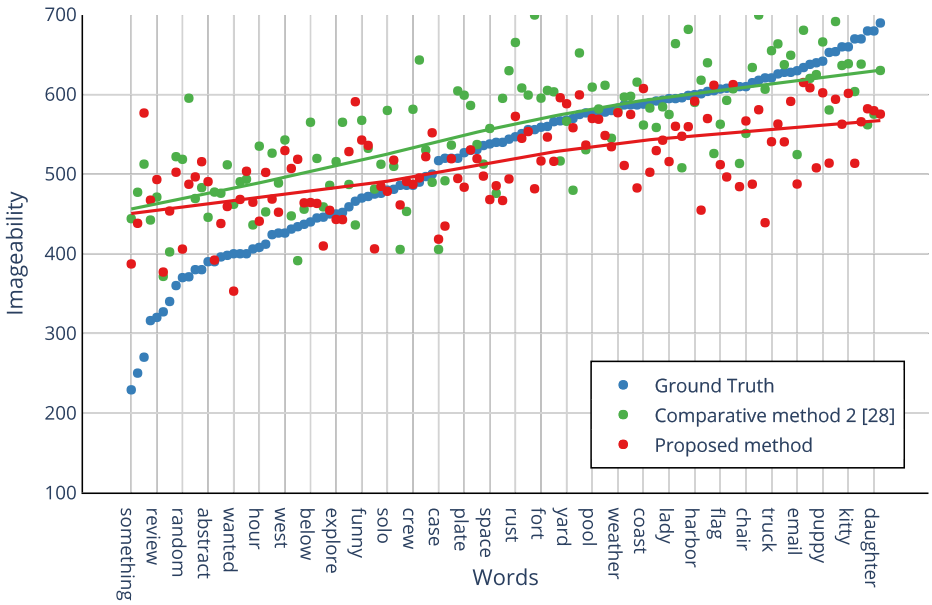

Fig. 5 Scatter plot of predicted values. To better understand the correlation between ground-truth values and predicted values, this scatter-plot shows predictions from the testing dataset. The labels on the horizontal axis is sampled across the dataset to get a feeling from which kind of word lies where on the spectrum. For comparison, values from the comparative method 2 (Text data mining) [28] are also plotted. All values are normalized to the range of [100, 700] to match the seven-level Lickert scale of the ground-truth data

5.1.4 Evaluation metrics

All experiments are evaluated using two metrics: First, the Mean Absolute Error (*MAE*) with the best result being 0 meaning no error compared to the ground-truth annotations. Second, the Pearson 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 values are rather close to their actual ground-truth value, even if they would result in the wrong ranking order due to slight differences. As the ground-truth seven-level Lickert scale 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.

5.2 Results

In the first experiment, the proposed method has been evaluated on the dataset of 587 words with 5,000 images each. Table 1 shows the results for each feature selection. It reaches

Table 1 Results for different sets of features

Feature	Correlation (1 : best)	MAE (0 : best)
L1: Color histograms	0.53	11.30
L2: SURF + Bag of Words	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	0.60	11.03
Combined High-level	0.61	10.18
Combined (Proposed method)	0.63	**10.14**
Comparative method 1 (Visual variety [24])	−0.01	67.31
Comparative method 2 (Text data mining [28])	**0.70**	10.39

While the ground-truth annotations are based on the Likert scale, the scores have been normalized to the range of [0,100] to improve understandability of the results. The dataset consists of 587 words and 5,000 images each

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 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, the method works similarly well for other datasets, where H1 features are not directly available.

For comparison, first, the algorithm used in our previous work [24] has been tested on the new dataset. In [24], the variety of visual characteristics in a BoW model is used to estimate a variety score for a dataset. It is closely related to the main assumption of this paper, although this previous work does not mention, nor evaluate, the possibility of imageability estimation. The result does not output, or have been trained with, imageability labels, but output a variety score based on relative variety differences between different datasets. Note that this previous work is largely a dataset-driven method, so the algorithm is left simple intentionally. The dataset used in this comparison has no relationship to the dataset-driven method proposed in [24], but is given to evaluate the performance of the algorithm. The results show, that the performance is vastly improved.

As a second comparative method, we compared our predicted imageability values to the method proposed by Ljubesic et al. [28] We used the pre-calculated dataset provided in their GitHub repository for the comparison. Note that these may not have used the identical ground-truth values for training, but the task of predicting imaegability for dictionary extension is the same. Their method predicts imageability entirely based on text data-mining, while ours exclusively uses visual characteristics of images. This makes for an interesting comparison between different modalities. The results show a slightly better correlation of

Table 2 Comparing dataset sizes

Dataset		Correlation (1 : best)	MAE (0 : best)
Fixed number of images	156 words trained with 5,000 images each	0.51	12.17
	312 words trained with 5,000 images each	0.62	10.58
	469 words trained with 5,000 images each	**0.63**	**10.14**
Fixed number of words	469 words trained with 1,000 images each	0.54	11.27
	469 words trained with 2,500 images each	0.57	10.87
	469 words trained with 5,000 images each	**0.63**	**10.14**

In this experiment, we evaluate how the number of words, or the number of images, relates to the overall performance of the approach. For each experiment, the combined feature set (Proposed method in Table 1) were used

0.70 for the text data mining method, but for the MAE, our 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 values using both textual and visual characteristics of multimodal datasets. However, due to the closely clustered 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, we 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 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, we trained the model with 312 (66% of training samples) and 156 (33% 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 our 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/word also further decreases, we did not look into larger dataset experiments.

In the third experiment, the regressor has been exchanged. We tried Random Forest, SVM, and a shallow Neural Network to determine which regression method works best on our data. As shown in Table 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

Table 3 Comparing 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 this experiment, we evaluate the chosen regressor used for the imageability prediction. For each experiment, we evaluated the proposed method using the combined feature set (Proposed method in Table 1) on a dataset with 587 words and 5,000 images each

469 training samples, we should keep in mind that the models are foremost training on the distribution of 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 we choose $n = 30$ conservatively to be on the secure side.

In the fourth experiment, the effect on different visual features on the imageability estimation for high- and low-imageability has been analyzed separately. As high-imageability and low-imageability words correlate with *concrete* words and *abstract* words, the visual characteristics of images in each word's dataset are very different. While high-imageability words share similar concrete objects or scenes, the low-imageability words have much more

Table 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 histograms	0.32	11.36	0.00	11.25
	L2: SURF/BoW	0.36	11.26	0.18	11.71
	L3: GIST	0.20	12.18	0.20	12.82
High-level	H1: Image theme	0.26	11.37	0.19	9.32
	H2: Image content	0.11	12.41	0.10	12.69
	H3: Image composition	−0.01	13.99	−0.05	13.87
Combined	Low-level features only	0.32	**10.90**	0.16	11.37
	High-level features only	0.27	11.31	0.10	9.10
	All (Proposed method)	0.26	10.79	**0.17**	10.11
Comparative	Text data mining [28]	**0.40**	13.27	0.18	**7.51**

The regressor is trained using the whole training data set. The testing samples were split in half around the imageability median. The upper half is considered concrete, while the lower half is considered abstract. The dataset consists of 587 words and 5,000 images each

noise and mostly share similar *atmosphere*, *backgrounds*, or the like. When splitting the testing dataset into two parts around the median imageability value of the ground-truth labels, 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. We can see that 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*).

The fifth experiment shows preliminary results for different parts-of-speech. Similar to the analysis of abstract vs. concrete words, we were 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 enough to evaluate the trends for each part-of-speech. The results in Table 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 1, the high-level feature H1 shows the best performance as a single feature. As the model for Table 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*

Table 5 Feature comparison for different parts-of-speech

Feature	Noun (32)		Adjective (33)		Adverb (18)		Verb (11)		Other (24)	
	Corr.	MAE	Corr.	MAE	Corr.	MAE	Corr.	MAE	Corr.	MAE
L1: Color	0.31	11.38	0.64	**14.45**	0.32	31.51	0.85	19.07	0.20	33.17
L2: SURF	0.35	11.02	0.27	18.32	0.14	33.35	**0.90**	20.35	0.27	31.45
L3: GIST	0.40	11.15	0.36	17.28	−0.02	32.33	0.89	20.27	0.43	**29.57**
H1: Theme	0.67	**8.69**	0.50	16.31	**0.56**	29.11	0.85	**17.71**	**0.85**	30.99
H2: Content	0.38	11.07	0.23	17.07	0.35	36.46	0.81	22.44	0.00	36.12
H3: Comp.	0.35	11.28	0.36	17.13	−0.10	37.83	0.56	25.91	0.28	34.06
Low-level	0.42	10.36	0.65	14.68	0.02	31.59	0.77	19.90	0.32	31.40
High-level	0.60	9.05	0.47	16.36	0.51	**28.31**	0.76	17.81	0.49	30.78
Proposed	0.65	9.17	0.53	15.42	0.29	29.08	0.79	18.13	0.60	30.47
Text [28]	**0.70**	10.36	**0.74**	13.63	0.25	34.81	0.63	22.69	0.39	33.25

As the obtained dataset predominantly consists of nouns, the regressor is trained using training dataset only containing nouns. This way, the testing dataset contains enough samples of non-nouns to show a meaningful analysis. The number of testing samples for each part-of-speech is given in brackets. The dataset consits of 587 words and 5,000 images each. The bold results indicate the best result per part-of-speech

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 6, which compares the ground-truth annotations to the predicted values for a selection of words. The three sections show words from the testing dataset, analyzing the results for high-imageability words, low-imageability words, and some outliers where the prediction failed, respectively. The examples show a close resemblance to the ground-truth values, 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 values off, preserving the general trend for most words.

To get a better understanding of the correlation between ground-truth values and the predicted values, Fig. 5 shows a scatter plot of the predicted testing dataset. Comparing the results of the proposed method with the comparative method, we can see that the global trend of each almost match exactly, but shifted along the vertical axis. Lowly-imageable words are such words that are thought to be harder to estimate due to their vagueness and abstractness. The scatter plot suggests that the proposed method works better towards lowly-imageable words, despite the bias of the training dataset, compared to the text data-mining method from [28]. Note that while the proposed method only uses 469 samples for training, the datasets used in [28] were in average about a magnitude larger.

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 5 are discussed,

Table 6 Prediction results of the proposed method

	Word	Predicted value	Ground-truth annotation
High-imageability	breakfast	591	628
	leaf	613	607
	plant	612	605
	coast	607	588
	pool	570	577
Low-imageability	early	390	391
	random	405	370
	challenge	438	396
	need	377	326
	break	459	397
Outliers (Worst 5)	fauna	577	270
	review	319	493
	silver	439	620
	email	487	630
	plastic	507	640

The predicted results and ground-truth annotations of a selection of high-imageability words, low-imageability words, and some outliers where the prediction failed, are shown. Either values are normalized to the interval of [100, 700] to match the seven-level Lickert scale of the ground-truth datasets

including the implications that visual feature selection might have for applications using imageability and multiple modalities.

6.1 Performance and feature selection

In the best feature selection, the proposed method yields a mean average error 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 a Lickert scale, the results in Table 6 are converted to the range of [100, 700] to match the ground-truth annotations. As shown, the error is smaller than one level on the Lickert scale, 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, our previous work [24] 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 for concrete words. This is unsurprising, 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. We initially expected that this would directly correlate to the performance for high-imageability words vs. low-imageability words. 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 image set. 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. 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 1, and comparatively lower correlation when evaluating only abstract or only concrete words (as shown in Table 4). This indicates that the general trend of high- vs. low-imageability words 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 annotations. When looking at the dataset, many concrete terms are clustered closely around the score 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 adjectives and adverbs seem intuitively 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.

6.2 Comparison to other methods

When comparing the results to our previous method [24], the proposed method shows a major improvement for the regression of imageability.

One thing worth mentioning is the goal of each paper. This paper focuses on imageability as a pre-existing concept, and proposes a method to regress given imageability scores from Psycholinguistic research as closely as possible. In contrast, the previous work was intended to measure slight differences of variety in direct comparison of words within the same domain. Therefore, the previous work excels in measuring the gap between *sports car*, *car*, *motor vehicle*, and *vehicle*. Such a detailed cross-comparison of words often makes less sense on the scale of a full dictionary, as a seven-level Lickert scale would not contain detail for this. Therefore, we regard both works as complementary approaches for different purposes.

When looking at the results shown in Table 1, it is clear that the previous method did not work for imageability. This has several reasons: First, recall that our previous method heavily relied on a modified custom dataset to create the *ideal* composition for further processing. This was unfeasible for a dataset with more than a few dozens of words, and is also complicated when crossing multiple domains. Next, the data-mining of the previous method relied on mean-shift clustering on top of a single visual feature. When analyzing words across different domains, it is hard to find a reference point for the number of clusters. While *car* has obviously more clusters than *sports car*, the relationship of clusters when comparing unrelated terms is not well defined. Thus, comparing the number of clusters between *car* and *pizza* will not give meaningful results. In contrast, the cross-similarity of images used for the approach proposed in this paper is a well-defined concept, even across different domains. On top of that, the variety of newly introduced visual features ensure a broader view on the data from additional angles. The results show that the features can complement each other, which is especially important as low-imageability words and high-imageability words show very different visual characteristics.

When comparing to Ljubešić et al. [28], 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. [28] has a better correlation. This suggests that the predicted labels 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 have 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, we believe that a correct order is vague and the general trend of predicted scores is the 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 low and mid-imageability terms. These results are probably also strengthened by the proposed method intrinsically focusing on noise analysis. Thus, concepts with a high visual variety usually being 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/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.

The text-only approach proposed by Ljubešić et al. [28] 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 would 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. As such, for future research, we plan to try a combined method analyzing both textual and visual features of co-existing text and images.

6.3 On the 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. [43] suggests that there is no upper limit for improving machine-learned models by increasing the amount of data, just a logarithmic diminishing return. Therefore, and due to the increased processing time, we have not further increased the number of images, although it can be assumed that the error can be decreased by further increasing the dataset.

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 image set. 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, we will summarize a few limitations and potential issues of the dataset creation process presented in this paper. The switch from a re-composited custom dataset in our previous work [24] 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 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 annotation, however, is rather abstract, as the term is usually associated with biology, making it a rather *hard* and *sciency* 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 midly imageable. In the ground-truth annotation, 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 image set for *craft* might be a mixture of *handcraft*, *aircraft*, and *watercraft* (which arguably makes the term rather abstract due to the ambiguity).

6.4 Applications

As discussed previously, the proposed method can be used to expand the vocabulary for imageability dictionaries for the general language. Due to the predicted labels in average rounding on the correct level on the Likert scale, we believe that the proposed method is sufficiently accurate for automatic creation of extended dictionaries. While a majority of existing datasets focus on nouns, the proposed method works for any type of part-of-speech. 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 entries in imageability dictionaries exist. While proper nouns are not commonly looked at in psycholinguistics, the possibility of creating ratings for such words would be beneficial for word selection tasks in image captioning and the similar. Lastly, it could also be extended to evaluations for other Psycholinguistic ratings, including sentiments.

When looking at image captioning [1], the results could be used to evaluate the understandability of generated text captions. Comparing text and image, concrete details and abstract concepts often supplement each other. An additional knowledge of this could yield a metric to assess the quality of auto-generated captions. Thus, in future research, it could be used to assess the accessibility, or the degree of information, in auto-generated texts.

As a metric, the measurements can be used for model understanding, which became important in recent years when considering the popularity of Neural Networks and the downsides of black-boxed methods. Furthermore, the results could be included in tools and datasets for NLP and sentiment research, like Empath [15], as they provide additional insight on semantic text understanding.

7 Conclusion

In this paper, we 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 small data samples compared to the full word corpora of languages.

The evaluations show a mean absolute error of 10.14 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. 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. An implementation and pre-trained model of the proposed method will be made available on GitHub[4]. 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.

For future research, it would be interesting to combine our analysis towards the analysis of visual characteristics for imageability estimation with existing methods using text data-mining for imageability and concreteness estimation to get a more round image of visual and textual meta-data contributing to the mental image of concepts. Furthermore, we want to expand the dataset furthermore. An analysis across multiple languages would be interesting, as imageability dictionaries already exist for other languages than English, like Indonesian [41], Cantonese [51], or Japanese. Lastly, we want to add other types of both low- and high-level visual features, to see if they can further improve the results.

References

1. Bai S, An S (2018) A survey on automatic image caption generation. Neurocomputing 311:291–304. https://doi.org/10.1016/j.neucom.2018.05.080
2. Balahur A, Mohammad SM, Hoste V, Klinger R (eds.) (2018) Proc. 9th Workshop on Computational Approaches to Subjectivity Sentiment and Social Media Analysis, ACL, Stroudsburg, PA, USA
3. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-Up Robust Features (SURF). Comput Vis Image Underst 110(3):346–359. https://doi.org/10.1016/j.cviu.2007.09.014
4. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
5. Charbonnier J, Wartena C (2019) Predicting word concreteness and imagery. In: Proc. 13th Int. Conf. on Computational Semantics, pp 176–187. https://www.aclweb.org/anthology/W19-0415
6. Chollet F, et al. (2015) Keras. https://github.com/fchollet/keras/
7. Coltheart M (1981) The MRC psycholinguistic database. Q J Exp Psychol A 33(4):497–505. https://doi.org/10.1080/14640748108400805
8. Coltheart V, Laxon VJ, Keating C (1988) Effects of word imageability and age of acquisition on children's reading. Br J Psychol 79(1):1–12. https://doi.org/10.1111/j.2044-8295.1988.tb02270.x

[4] https://github.com/mkasu/imageabilityestimation/

9. Comaniciu D, Meer P (2002) Mean Shift: A robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24(5):603–619. https://doi.org/10.1109/34.1000236
10. Cortese MJ, Fugett A (2004) Imageability ratings for 3,000 monosyllabic words. Behav Res Methods Instrum Comput 36(3):384–387. https://doi.org/10.3758/BF03195585
11. Csurka G, Dance CR, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Proc. ECCV 2004 Workshop on Statistical Learning in Computer Vision, pp 1–22
12. Deng JDJ, Dong WDW, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: Proc. 2009 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, pp 2–9, https://doi.org/10.1109/CVPR.2009.5206848
13. Divvala SK, Farhadi A, Guestrin C (2014) Learning everything about anything: Webly-supervised visual concept learning. In: Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition, pp 3270–3277, https://doi.org/10.1109/CVPR.2014.412
14. Douze M, Jégou H, Sandhawalia H, Amsaleg L, Schmid C (2009) Evaluation of GIST descriptors for Web-scale image search. In: Proc. ACM Int. Conf. on Image and Video Retrieval 2009, pp 19:1–19:8. https://doi.org/10.1145/1646396.1646421
15. Fast E, Chen B, Bernstein MS (2016) Empath: Understanding topic signals in large-scale text. Computing Research Repository. arXiv:1602.06979
16. Giesbrecht B, Camblin CC, Swaab TY (2004) Separable effects of semantic priming and imageability on word processing in human cortex. Cereb Cortex 14(5):521–529
17. Hessel J, Mimno D, Lee L (2018) 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, pp 2194–2205. https://doi.org/10.18653/v1/N18-1199
18. Hewitt J, Ippolito D, Callahan B, Kriz R, Wijaya DT, Callison-Burch C (2018) Learning translations via images with a massively multilingual image dataset. In: Proc. 56th Annual Meeting of the Association for Computational Linguistics, vol 1, pp 2566–2576. https://doi.org/10.18653/v1/P18-1239
19. Holzinger A, Biemann C, Pattichis CS, Kell DB (2017a) What do we need to build explainable AI systems for the medical domain. Computing Research Repository. arXiv:1712.09923
20. Holzinger A, Malle B, Kieseberg P, Roth PM, Müller H, Reihs R, Zatloukal K (2017b) Towards the augmented pathologist: Challenges of explainable-AI in digital pathology. Computing Research Repository. arXiv:1712.06657
21. Inoue N, Shinoda K (2016) Adaptation of word vectors using tree structure for visual semantics. In: Proc. 24th ACM Multimedia Conf., pp 277–281. https://doi.org/10.1145/2964284.2967226
22. Itseez (2015) Open source computer vision library. https://opencv.org/
23. Jones GV (1985) Deep dyslexia, imageability, and ease of predication. Brain Lang 24(1):1–19. https://doi.org/10.1016/0093-934X(85)90094-X
24. Kastner MA, Ide I, Kawanishi Y, Hirayama T, Deguchi D, Murase H (2019) Estimating the visual variety of concepts by referring to Web popularity. Multimed Tools Appl 78(7):9463–9488. https://doi.org/10.1007/s11042-018-6528-x
25. Kawakubo H, Akima Y, Yanai K (2010) Automatic construction of a folksonomy-based visual ontology. In: Proc. 2010 IEEE Int. Symposium on Multimedia, pp 330–335. https://doi.org/10.1109/ISM.2010.57
26. Kohara Y, Yanai K (2013) Visual analysis of tag co-occurrence on nouns and adjectives. In: Li S, El Saddik A, Wang M, Mei T, Sebe N, Yan S, Hong R, Gurrin C (eds) Advances in Multimedia Modeling: 19th Int. Conf. on Multimedia Modeling Procs., Springer, Lecture Notes in Computer Science, vol 7732, pp 47–57. https://doi.org/10.1007/978-3-642-35725-1-5
27. Li JJ, Nenkova A (2015) Fast and accurate prediction of sentence specificity. In: Proc. 29th AAAI Conf. on Artificial Intelligence, pp 2281–2287
28. LjubeŠić N, FiŠer D, Peti-Stantić A (2018) Predicting concreteness and imageability of words within and across languages via word embeddings. In: Proc. 3rd Workshop on Representation Learning for NLP, pp 217–222, https://doi.org/10.18653/v1/W18-3028
29. Loper E, Bird S (2002) 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, pp 63–70. https://doi.org/10.3115/1118108.1118117
30. Ma W, Golinkoff RM, Hirsh-Pasek K, McDonough C, Tardif T (2009) Imageability predicts the age of acquisition of verbs in Chinese children. J Child Lang 36:405–423. https://doi.org/10.1017/S0305000908009008
31. Miller GA (1995) WordNet: A lexical database for English. Comm ACM 38(11):39–41. https://doi.org/10.1145/219717.219748
32. Paivio A, Yuille JC, Madigan SA (1968) Concreteness, imagery, and meaningfulness values for 925 nouns. J Exp Psychol 76(1):1–25

33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
34. Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic Inquiry and Word Count: LIWC 2001. Erlbaum, Mahwah, NJ, USA
35. Redmon J, Farhadi A (2016) YOLO9000: Better, faster, stronger. Computing Research Repository. arXiv:1612.08242
36. Reilly J, Kean J (2010) Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. J Cogn Sci 31(1):157–168. https://doi.org/10.1080/03640210709336988
37. Ringeval F, Schuller B, Valstar M, Cowie R, Kaya H, Schmitt M, Amiriparian S, Cummins N, Lalanne D, Michaud A, Ciftçi E, Güleç H, Salah AA (2018) Proc. 2018 Audio/Visual Emotion Challenge and Workshop. ACM, New York, NY, USA, Pantic M (ed)
38. Samek W, Wiegand T, Müller K (2017) Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. Computing Research Repository. arXiv:1708.08296
39. Schwanenflugel PJ (2013) Why are abstract concepts hard to understand? in: The Psychology of Word Meanings, Psychology Press, New York, NY, USA, pp 235–262
40. Shu X, Qi GJ, Tang J, Wang J (2015) Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation. In: Proc. 23rd ACM Multimedia Conf., pp 35–44. https://doi.org/10.1145/2733373.2806216
41. Sianipar A, van Groenestijn P, Dijkstra T (2016) Affective meaning, concreteness, and subjective frequency norms for Indonesian words. Front Psychol 7:1907. https://doi.org/10.3389/fpsyg.2016.01907
42. Smolik F, Kriz A (2015) The power of imageability: How the acquisition of inflected forms is facilitated in highly imageable verbs and nouns in Czech children. J First Lang 35(6):446–465. https://doi.org/10.1177/0142723715609228
43. Sun C, Shrivastava A, Singh S, Gupta A (2017) Revisiting unreasonable effectiveness of data in deep learning era. Computing Research Repository. arXiv:1707.02968
44. Tanaka S, Jatowt A, Kato MP, Tanaka K (2013) Estimating content concreteness for finding comprehensible documents. In: Proc. 6th ACM Int. Conf. on Web Search and Data Mining, pp 475–484. https://doi.org/10.1145/2433396.2433455
45. Tang J, Shu X, Li Z, Qi GJ, Wang J (2016) Generalized Deep Transfer Networks for Knowledge Propagation in Heterogeneous Domains. ACM Trans. Multimed Comput. Commun. Appl. 12(4s):1–22. https://doi.org/10.1145/2998574
46. Tang J, Shu X, Qi G, Li Z, Wang M, Yan S, Jain R (2017) Tri-clustered tensor completion for social-aware image tag refinement. IEEE Trans Pattern Anal Mach Intell 39(8):1662–1674. https://doi.org/10.1109/TPAMI.2016.2608882x
47. Tang J, Shu X, Li Z, Jiang Y, Tian Q (2019) Social anchor-unit graph regularized tensor completion for large-scale image retagging. IEEE Trans Pattern Anal Mach Intell 41(8):2027–2034. https://doi.org/10.1109/TPAMI.2019.2906603
48. Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li LJ (2016) YFCC100M: The new data in multimedia research. Comm ACM 59(2):64–73. https://doi.org/10.1145/2812802
49. Vidanapathirana M (2018) YOLO3-4-Py. https://github.com/madhawav/YOLO3-4-Py
50. Yanai K, Barnard K (2005) Image region entropy: A measure of "visualness" of Web images associated with one concept. In: Proc. 13th ACM Multimedia Conf., pp 419–422. https://doi.org/10.1145/1101149.1101241
51. Yee LT (2017) Valence, arousal, familiarity, concreteness, and imageability ratings for 292 two-character Chinese nouns in Cantonese speakers in Hong Kong. PloS one 12(3):e0174569. https://doi.org/10.3389/fpsyg.2016.01907
52. Zhang M, Hwa R, Kovashka A (2018) Equal but not the same: Understanding the implicit relationship between persuasive images and text. In: Proc. British Machine Vision Conference 2018, no. 8

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

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

Frank Nack is tenure Associate Professor at the INtelligent Data Engineering Lab (INDE lab) of the Informatics Institute at the University of Amsterdam (UvA). The main thrust of his research is on context and process aware media knowledge spaces (design, interaction, memory); interactive storytelling, the representation and adaptation of experiences; the representation and support of creativity in the domain of arts and entertainment; user modelling, computational applications of media theory and semiotics; and computational humour theory. He published more than 100 papers on these topics. He has been serving on program committees (member or chair) at the major conferences: ACM MM, ACM Hypertext, Multimedia Modeling (MMM), ICME, and ICIDS. He frequently reviews for ACM TOMCCAP, Multimedia Tools and Applications, The Information Retrieval Journal, IEEE Intelligent Systems Magazine, IEEE Multimedia, International Journal of Continuing Engineering Education and Life Long Learning, New Review of Hypermedia and Multimedia. He is member of the ACM, ACM SIGMM, ACM SIGCHI and ACM SIGWEB.

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

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

Daisuke Deguchi received his BEng and MEng in Engineering and PhD in Information Science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively. He is currently an Associate Professor in Information Strategy Office, Nagoya University, Japan. He is working on object detection, segmentation, and recognition from videos, and their applications to ITS technologies, such as detection and recognition of traffic signs.

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

Affiliations

Marc A. Kastner[1] · Ichiro Ide[1] · Frank Nack[2] · Yasutomo Kawanishi[1] · Takatsugu Hirayama[3] · Daisuke Deguchi[4] · Hiroshi Murase[1]

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

Frank Nack
nack@uva.nl

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

Takatsugu Hirayama
takatsugu.hirayama@nagoya-u.jp

Daisuke Deguchi
ddeguchi@nagoya-u.jp

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

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

2 Informatics Institute, University of Amsterdam, Amsterdam, 1098 XH, The Netherlands

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

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