Imageability Estimation using Visual and Language Features

Chihaya Matsuhira
matsuhirac@murase.is.i.nagoya-u.ac.jp
School of Engineering,
Nagoya University
Nagoya, Aichi, Japan

Marc A. Kastner
kastnerm@murase.is.i.nagoya-u.ac.jp
Graduate School of Informatics,
Nagoya University
Nagoya, Aichi, Japan

Ichiro Ide
ide@i.nagoya-u.ac.jp
Mathematical & Data Science Center,
Nagoya University
Nagoya, Aichi, Japan

Yasutomo Kawanishi
kawanishi@i.nagoya-u.ac.jp
Graduate School of Informatics,
Nagoya University
Nagoya, Aichi, Japan

Takatsugu Hirayama
takatsugu.hirayama@nagoya-u.jp
Institute of Innovation for Future
Society, Nagoya University
Nagoya, Aichi, Japan

Keisuke Doman
kdoman@sist.chukyo-u.ac.jp
School of Engineering,
Chukyo University
Toyota, Aichi, Japan

Daisuke Deguchi
ddeguchi@nagoya-u.jp
Graduate School of Informatics,
Nagoya University
Nagoya, Aichi, Japan

Hiroshi Murase
murase@i.nagoya-u.ac.jp
Graduate School of Informatics,
Nagoya University
Nagoya, Aichi, Japan

ABSTRACT

Imageability is a concept from Psycholinguistics quantizing the human perception of words. However, existing datasets are created through subjective experiments and are thus very small. Therefore, methods to automatically estimate the imageability can be helpful. For an accurate automatic imageability estimation, we extend the idea of a psychological hypothesis called *Dual-Coding Theory*, that discusses the connection of our perception towards visual information and language information, and also focus on the relationship between the pronunciation of a word and its imageability. In this research, we propose a method to estimate imageability of words using both visual and language features extracted from corresponding data. For the estimation, we use visual features extracted from low- and high-level image features, and language features extracted from textual features and phonetic features of words. Evaluations show that our proposed method can estimate imageability more accurately than comparative methods, implying the contribution of each feature to the imageability.

CCS CONCEPTS

• **Information systems** → **Sentiment analysis**; • **Applied computing** → **Psychology**; • **Computing methodologies** → *Language resources*; Phonology / morphology.

KEYWORDS

Multimedia modeling, language and vision, computational psycholinguistics

ACM Reference Format:
Chihaya Matsuhira, Marc A. Kastner, Ichiro Ide, Yasutomo Kawanishi, Takatsugu Hirayama, Keisuke Doman, Daisuke Deguchi, and Hiroshi Murase. 2020. Imageability Estimation using Visual and Language Features. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372278.3390731

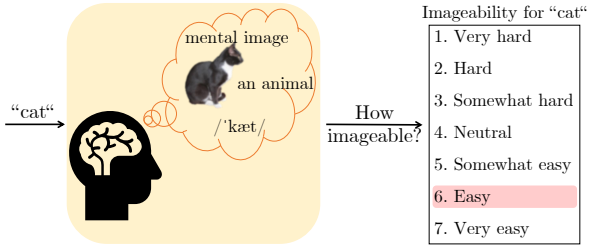

Figure 1: Process of determining *Imageability*. Given a word, human subjects guess how imageable the word is, imaging various features of the word such as its mental images, meanings, and pronunciations.

1 INTRODUCTION

Imageability is a concept proposed by Paivio et al. [9] in Psycholinguistics as a measurement for quantizing the human perception of words. It explains how easily a person can form a mental image associated to a word. The imageability of a word would be high when its concept is easy to imagine, and low when its concept is hard to imagine. Therefore, we can say that the imageability represents how humans perceive the world.

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3390731

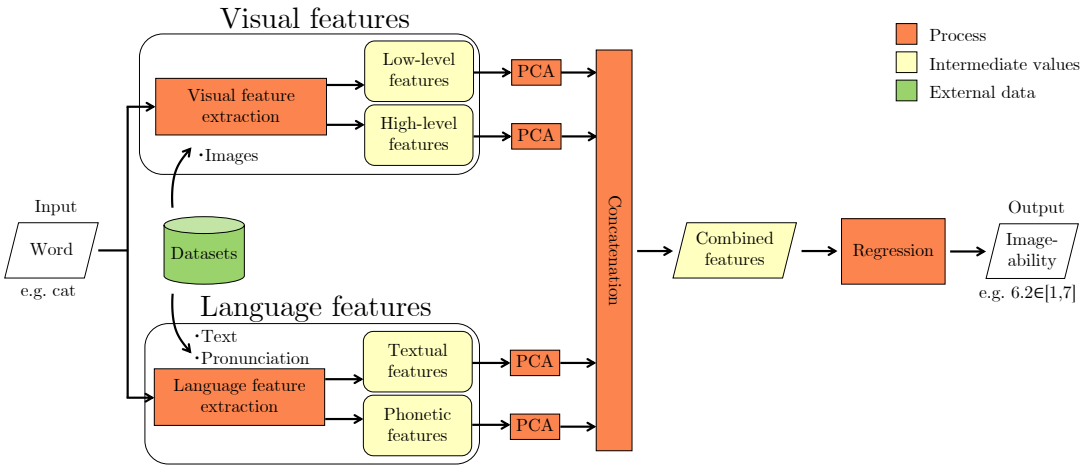

Figure 2: Flowchart of the proposed method. Given an input word, visual features and language features are extracted separately. PCA is applied to each feature to uniform the dimensionality. The transformed features are then concatenated and used to train a Random Forest regressor to predict imageability scores.

Although the use of imageability in multimedia applications has not been thoroughly researched yet, existing research on multimodal data analysis [13] showcases some first use-cases with promising results for a better understanding of semantics. Improvements of image retrieval and image captioning could be considered as other promising use-cases.

Datasets used in this field are created in very laborious experiments. (Figure 1 shows an example of the process.) However, because of the massive vocabulary of natural language including new words, obtaining imageability scores of all words through crowd-sourcing is not feasible. Therefore, a method to automatically estimate the score for each word should be helpful to extend the vocabulary of existing dictionaries.

In the field of Informatics, there are several pieces of research on the estimation of imageability scores of words using data-mining. In the area of Computer Vision (CV), Kastner et al. [3] proposed a method to estimate imageability scores using Web-crawled images. They assumed an intrinsic relationship between the imageability of words; human perception of our environment, and contents of image data we upload onto Social Media. They focused on the cross-similarity between images for each word and estimated imageability scores based on visual features extracted from those images. On the other hand, in the area of Natural Language Processing (NLP), Ljubešić et al. [4] estimated imageability and concreteness scores exploiting word-embeddings like the pretrained fastText [1] to estimate these scores within a single or across multiple languages. Although these methods exist, they do not use both visual information and language information simultaneously.

To estimate imageability scores more accurately, we utilize a well-known psychological hypothesis called "Dual-Coding Theory" [8]. This postulates that we humans process visual information and language information separately for encoding information. Based on this, we aim to improve the accuracy of the imageability estimation by using these two types of information.

Moreover, we make an assumption that we unconsciously use information about how we pronounce a word when imaging the word. Hence we use pronunciation information as one of the language features of words.

In this research, based on the theory and the assumption above, we propose a method to estimate the imageability scores of words using both visual features and language features. As the visual features, we extract a variety of low- and high-level image features following Kastner et al. [3]. As the language features, we take advantage of word-embeddings and phonetic information of words.

In Section 2, previous research related to the imageability estimation is discussed. Section 3 proposes our method for automatically estimating the imageability of words via extracting visual, textual, and phonetic features. Lastly, Section 4 evaluates the method with an experiment, discussing the results in comparison with existing methods before concluding the paper in Section 5.

2 METHOD

In this research, we propose a method to estimate the imageability of words via visual and language feature mining. The proposed method consists of two steps: First, visual features and language features are extracted separately from different sources of information (datasets). Second, these feature vectors are merged, and the imageability score for a word is estimated with a regressor. The flowchart of the proposed method is shown in Figure 2.

2.1 Visual Feature Extraction

The visual feature extraction is based on Kastner et al.'s method [3]. First, 5,000 images for an input word are crawled from Social Media. Then, various types of visual features are extracted for each image. These visual features are separated into low-level and high-level visual features, according to the visual information they contain. The low-level visual features encode information about colors and gradients of an image, while the high-level features encode information about what and where objects are in an image.

For each type of feature, a similarity matrix is calculated among images crawled for each word to encode the cross-similarity among them. Lastly, a combined vector of the largest 30 Eigenvalues of the similarity matrix is extracted as the similarity vector.

The final low- and high-level visual features are obtained by concatenating these similarity vectors respectively.

2.2 Language Feature Extraction

Two types of language features are adopted: A textual feature and two phonetic features. According to the result of Ljubešić et al. [4], textual features are proved to be useful for imageability estimation. Additionally, assuming that pronunciations of words would have an influence on imageability scores, we adopt phonetic features.

As the textual feature, a word-embedding extracted with pretrained models like word2vec [5][7] or fastText [1] is used. Such a word-embedding is presumed to contain information on the position of a word in a sentence and its word co-occurrences, and has previously been shown to work well by Ljubešić et al. [4]. For this reason, we consider using this word-embedding as a source of information for imageability estimation.

The phonetic features are extracted from pronunciation information of each word. In this research, we represent the pronunciation of words with the International Phonetic Alphabet (IPA)[1]. To extract the phonetic features from this, we propose two methods that are analogies of existing text-feature extraction methods.

First, we propose the Bag of Phonemes (BoP) feature. In this method, based on the Bag of Words (BoW) algorithm, any occurrence of phoneme in an input pronunciation is counted to make a multiplicity vector. During this process, when a primary stress character appears, the following vowel is weighted with $k = 3$.

The second is the pron2vec (pronunciation to vector) feature. Here, expanding the idea of a simple word2vec-LSTM model for a sentence embedding, we train a model that receives the pronunciation of a word as an input and outputs an embedding vector. In this method, first, the pron2vec model is pretrained with imageability scores. Finally, the output of the intermediate-layer of pron2vec is extracted as a feature.

In the proposed method, both BoP and pron2vec features are used as phonetic features. The overall extraction process for these phonetic features extraction is shown in Figure 3. Note that for both features, the phonemes are restricted to only vowels, consonants, and the primary stress. All other phonemes and auxiliary symbols are skipped.

2.3 Merging and Training

For each feature extracted; low-level visual features, high-level visual features, textual features, BoP phonetic features, and pron2vec features, Principal Component Analysis (PCA) is applied in order to unify the dimensionality of each feature vector. After that, these feature vectors are concatenated. Lastly, a Random Forest model is trained on the concatenated feature vector to regress an estimated imageability score.

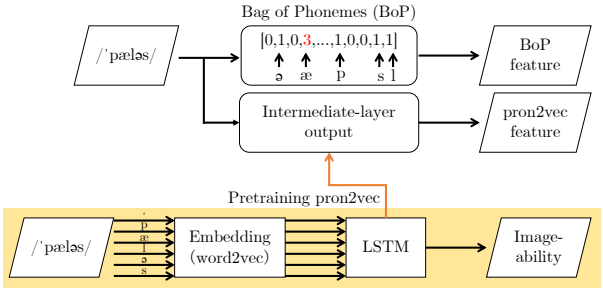

Figure 3: Extraction of phonetic features. Two types of features are extracted separately as BoP and pron2vec. The BoP features count the occurrence of phonemes. For the extraction of the pron2vec features, first the whole LSTM model is pretrained with imageability, using an already pretrained word2vec model as an embedding. The intermediate layer output of this LSTM model is extracted as the pron2vec features.

3 EVALUATION

For the evaluation of the proposed method, an experiment was conducted. First, we will discuss the actual implementation of our experiments, as well as our datasets. Second, we will report the results, and discuss findings and observations.

3.1 Implementation

For visual features, three low-level and three high-level features were extracted. Low-level visual features were composed of HSV color histogram, GIST features, and SURF features. High-level visual features were composed of Visual Concept, image contents, and composition of an image. The Visual Concept was derived from autotags attached to each image in the used YFCC100M dataset [12], and both the content and the composition of an image were calculated via object detection performed by YOLO9000 [10]. The textual features were extracted with the pretrained fastText model published by Facebook [6].

For comparative methods, we cited the result calculated by Kastner et al. [3] and Ljubešić et al. [4]. We also tested all possible combinations of the visual and language features of the proposed method. The difference between our single feature evaluations and the previous methods is that the PCA is applied in our method.

3.2 Dataset

For this experiment, we prepared data for 587 English words. These words were split into 469 for training and 118 for testing.

In this research, three types of datasets were used: an imageability dictionary for training the regressor and pron2vec, an image dataset, and a pronunciation dictionary.

First, for training both models, two English imageability dictionaries [2][11] were used. The imageability labels were stored as a value within a range of [1, 7] based on a seven-level Likert scale.

Second, for the image dataset, we used images from Flickr[2], obtained from the YFCC100M dataset [12] with about 1 million

[1] http://www.internationalphoneticassociation.org/content/ipa-chart/

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

Table 1: Experiment results. While the ground-truth labels are based on the Likert scale, the labels have been normalized to a range of [0,100] to improve the understandability of the results.

Features	MAE	Correlation
VIS	10.51	0.60
TXT	8.75	0.75
PRN	13.65	0.30
VIS + TXT	8.67	0.75
VIS + PRN	10.49	0.61
TXT + PRN	8.56	0.75
ALL (Proposed: VIS + TXT + PRN)	**8.39**	**0.77**
Comparative 1 (Kastner et al. [3])	10.14	0.63
Comparative 2 (Ljubešić et al. [4])	10.39	0.70

images and videos from Flickr. Based on the meta-data attached to each image like title and tags, 5,000 images for each word were selected for the extraction of visual features.

Third, we used Macmillan Dictionary[3] as a source of American-English pronunciation for English words. We preprocessed the pronunciation data for single-syllable words having no stress character, so that those words would get a primary stress character in the front of their pronunciation.

3.3 Results

The overall results are shown in Table 1. Our proposed method using all visual and language features gave both the smallest Mean Average Error (MAE), and the best correlation to the ground-truth. Compared to the previous methods, our proposed method improved by at least 17% MAE (compared to [3]) and 10% correlation (compared to [4]). Moreover, we confirmed that in any combination of features, combining features contributed to the improvement of the accuracy of estimating imageability scores. From these results, we can infer that all three types of information; visual features, textual features, and phonetic features, complement each other.

Compared to the result of Kastner et al. [3], however, our method with only visual features worsened both metrics. This can be due to the application of PCA in our method, where both levels of visual features were transformed into the same dimensionality. In reality, the appropriate dimension for each feature may differ. Therefore, this process might have caused slight deterioration of the features.

In contrast, compared to the result of Ljubešić et al. [4], our method with only textual feature improved both metrics. In fact, our method was not a direct implementation of theirs, which has a focus on training imageability values across languages. Rather than that, we used a fastText pretrained on the English language, further post-processed with PCA to fit the dimensionality of the other feature spaces, and thus the result showed improvement.

Lastly, Figure 4 shows a scatter plot of the imageability scores estimated by the proposed method using single features and all the features for every word. The arranged lines shown are the Least Squares Regression Line (LSRL) for each method. We can see that the result of the proposed method resembles that of the method using only textual features, and due to the influence by the other two features, fits better to the ground-truth.

[3]https://www.macmillandictionary.com/

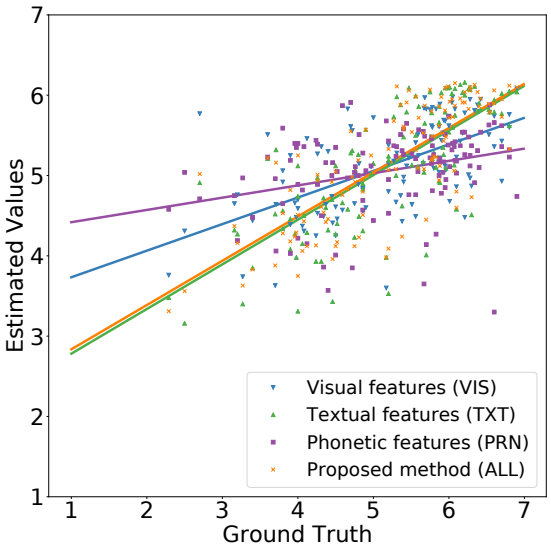

Figure 4: Scatter plot and LSRL of the predicted imageability scores. All scores are normalized to the range of [1, 7] to match the seven-level Likert scale of the ground-truth data.

4 CONCLUSION

In this research, we proposed a method to automatically estimate imageability scores of words using both visual and language features related to those words. The evaluation shows our proposed method improves the estimation by at least 17% MAE (compared to [3]) and 10% correlation (compared to [4]) compared to the comparative methods. This is considered to be resulting from the fact that visual and language features contain different aspects of information.

So far, methods to estimate imageability scores have been researched in different areas: CV and NLP. Our proposed method successfully merged these research progresses resulting in more accurate imageability estimation to the existing dataset. It can be used to enrich automatically created imageability datasets by providing predictions much closer to our perception.

In futere work, we plan to improve the overall model structure, including feature-merging procedures and the whole structure of our pron2vec model. Especially for the latter, currently we need to train the model with imageability scores, but this could not be the best approach since the process seems redundant. Hence, we are looking at using LSTM-AutoEncoder to extract phonetic features without referring to existing labels.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5 (6 2017), 135–146.

[2] Michael J Cortese and April Fugett. 2004. Imageability ratings for 3,000 monosyllabic words. *Beh. Res. Methods Instrum. Comput.* 36, 3 (8 2004), 384–387.

[3] Marc A. Kastner, Ichiro Ide, Frank Nack, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, and Hiroshi Murase. 2020. Estimating the imageability of words by mining visual characteristics from crawled image data. *Multimed. Tools Appl.* (2 2020). https://doi.org/10.1007/s11042-019-08571-4

[4] Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proc. 56th Annu. Meet. Assoc. Comput. Linguist.* 217–222.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Comput. Res. Repos., arXiv: 1301.3781.

[6] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proc. Int. Conf. on Language Resources and Evaluation 2018.* 52–55.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, Vol. 2. 3111–3119.

[8] Allan Paivio. 1990. *Mental representations: A dual coding approach.* Vol. 9. Oxford University Press, Oxford, England.

[9] Allan Paivio, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *J. Exp. Psychol.* 76, 1 (1 1968), 1–25.

[10] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, faster, stronger. In *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit.* 6517–6525.

[11] Jamie Reilly and Jacob Kean. 2007. Formal distinctiveness of high-and low-imageability nouns: Analyses and theoretical implications. *J. Cogn. Sci.* 31, 1 (2 2007), 157–168.

[12] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Comm. ACM* 59, 2 (2 2016), 64–73.

[13] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proc. 2018 Br. Mach. Vis. Conf.* 14p.