On Understanding Visual Relationships of Concepts by Visualizing Bag-of-Visual-Words Models

Marc A. KASTNER[1,a) Ichiro IDE[1,b) Yasutomo KAWANISHI[1,c) Takatsugu HIRAYAMA[1,d)
Daisuke DEGUCHI[1,e) Hiroshi MURASE[1,f)

Abstract

Recent applications in image processing often use a multimodal approach using both text and imagery. This is prone to semantic gap issues when converting between image and language. There has been few research quantifiying visual differences when assessing semantic relationships. In this research, we analyze datasets composed of logically related concepts. By visualizing a Bag-of-Visual-Words (BoVW) model spatially, visual semantics of logically related sub-concepts are shown. To find hidden semantics of related concepts, the most common visual words of an image in relation to its neighbors are highlighted. This provides additional semantic knowledge on how sub-ordinate concepts visually relate to another. It is thought to give an insight on the human perception of these concepts, and can be used in future research to estimate psycholinguistic ratings.

1. Introduction

In recent multimedia applications, approaches involving text, image, and video content is often used to combine knowledge spanning multiple modalities. The so-called semantic gap describes a number of problems that occur when transfering between modalities. Visual semantics can give an insight on human perception of given concepts. For example, largely overlapping sub-concepts might be less distinguishable, if they are also visually equal. In contrast, two very related concepts are more easily distinguishable, if visually distinct, even if they logically belong together. In psycholinguistics, these properties are called imagability and concreteness [10]. A quantification of this would greatly benefit word selection problems in various applications.

In this research, we visually analyze datasets composed of logically related concepts. A dataset is created by combining images from ImageNet [5] using the WordNet hierarchy [8]. A separate Bag-of-Visual-Words (BoVW) model is trained for each concept, using images of all its subordinate concepts. The model will prioritize keypoints standing out when visually comparing different concepts. By visualizing the resulting feature space spatially, hidden visual semantics of logically-related sub-concepts are shown. To aid in finding hidden semantics of related concepts, the most common visual words of an image in relation to its neighbors are highlighted. This provides an additional semantic knowledge on how sub-ordinate concepts visually relate to another, laying the ground work to estimate psycholinguistic ratings like imagability and concreteness.

Section 2 gives a brief overview of related work. In Section 3, the proposed idea is discussed. First, the creation of dataset and visual model are described in detail. Then, the approach to highlight important visual words as seen by the machine is outlined. Section 4 showcases our interactive tool, discussing possible gains in semantic knowledge through it.

2. Related Work

Research on how language interacts with human perception has been part of psycholinguistics. Paivio et al. [9] analyzed the concreteness, imagery, and meaningfulness of nouns. In the MRC Psycholinguistic Database by Wilson et al. [11], words are rated by familiarity, concreteness, imagability, and meaningfulness. More recent research by Cortese et al. [4] classifies imageability ratings for 3,000 words, which is thought to be useful for human

[1] Nagoya University
[a) kastnerm@murase.is.i.nagoya-u.ac.jp
[b) ide@i.nagoya-u.ac.jp
[c) kawanishi@i.nagoya-u.ac.jp
[d) takatsugu.hirayama@nagoya-u.jp
[e) ddeguchi@nagoya-u.jp
[f) murase@i.nagoya-u.ac.jp

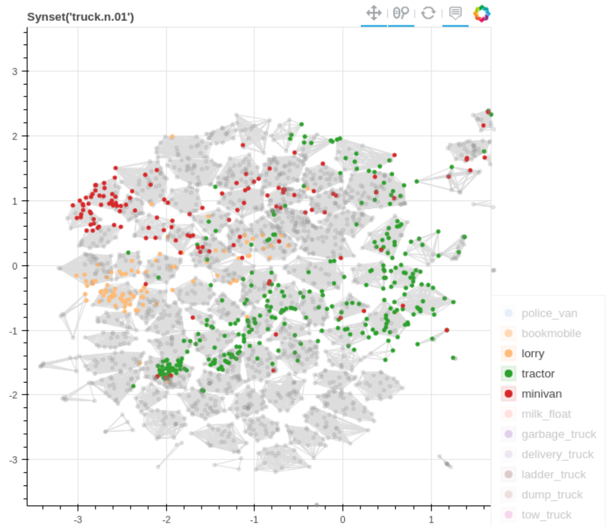


Fig. 1 Example of a visualized synset. Each sample corresponds to a single image within the imageset. The spatial distribution is based on a UMAP embedding of the BoVW model. The grey clusters shown are results of mean-shift clustering of visually similar neighbors. The labels visually highlight the subconcept affiliation of each image based on the WordNet hierarchy.

word recognition and memory studies.

In our previous work [13], we quantified the visual variety of concepts using image data. The evaluation verified the estimate to match human expectations by comparing it to the results of a user study.

There has been research on visualizing visual feature spaces, as it is of interest to understand how well object recognition works. Yue et al. [12] visualize the contents of a BoVW model. Using a modified model, they reverse the encoding and reconstruct images from the visual model. With this, they can visualize the degree of information lost in the representation. Hentschel et al. [6] use an object recognition classifier to visualize which regions of an image most likely contain the trained object. For a given image of an object, they create a probability heatmap highlighting which regions of the image most likely contain the object. Both projects use a feature visualization to judge the quality of the visual feature representations. There has not been research analyzing the semantic implications of such visualizations.

3. Approach

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

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

3.1 Dataset

To analyze visual relationships within concepts, a dataset which has a strong variety of subordinate concept images is needed. For each abstract synset, a set of related sub-concepts is generated by crawling its most subordinate concepts in the WordNet hierarchy. The most subordinate concepts in the WordNet graph are the leaf nodes below the abstract synset. Then, an image set is generated using ImageNet [5] images as a baseline. Instead of using the image sets provided by ImageNet directly, the images of its sub-concepts are merged. This is thought to provide a dataset with a higher variety, and thus preserving knowledge about hidden concept semantics. The information which image belongs to which sub-concept is preserved for labelling.

3.2 Visual representations

As a visual representation, a BoVW model is generated for each abstract synset separately. It is trained using images of its subordinate concepts using the previously created imagesets. For each image, visual features in form of Speeded Up Robust Features (SURF) [1] are used.

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

3.3 Visualization

For visualization purposes, Uniform Manifold Approximation and Projection (UMAP) [7] is used to compute a dimensionality reduced spatial embedding of the visual model. This embedding gives insight on the spatial distribution of different subordinate concepts within the visual feature space. Next, the goal is to highlight the most common visual words as seen by the machine, in relation to neighboring images. This allows to infer what the computer perceives as visually related parts of neighboring images. The process of selecting the most common visual

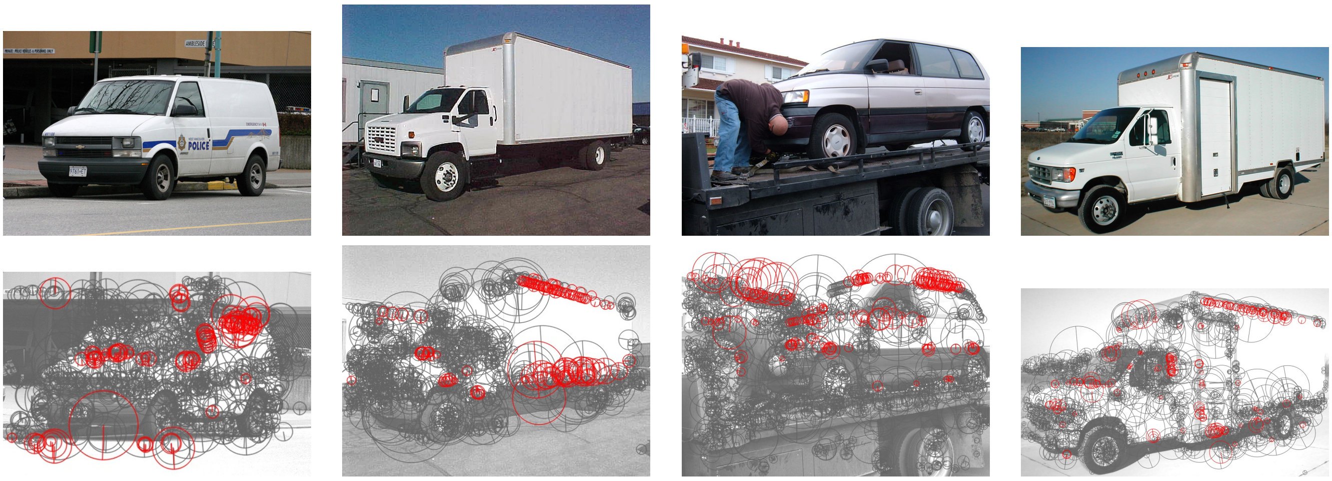


Fig. 2 Example of the common keypoint visualization. Four neighboring images within the dataset `truck`, which show common visual characteristics. The red keypoints are the 10% which share the most common visual words between the cluster of neighboring images.

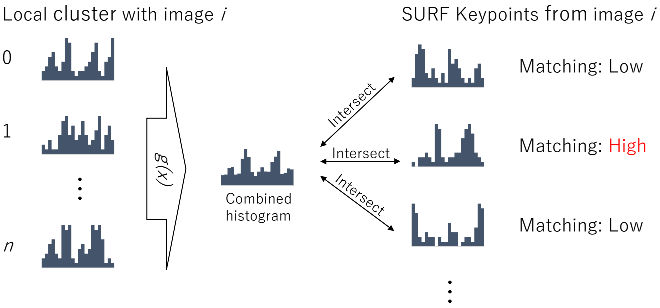


Fig. 3 Process of selecting common visual words. For each image i, the BoVW histograms of the neighboring images in its local cluster are combined using equation $g(x)$. Then, the resulting histogram is intersected with the BoVW histograms of each of its keypoints. The closest matching keypoints are chosen for visualization.

words for each image is shown in Fig. 3. For each image, a number of visually similar images are selected using Mean-Shift Clustering [3]. Then, the BoVW histograms $f(x)$ of all selected n images are merged using this equation:

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

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

Figure 2 shows an example of four neighboring images in an imageset within the imageset for the synset `truck`. While they belong to different subordinate concepts, they share visual similarities and are thus clustered together. The red regions in the bottom row highlight the most common visual words. As all images are shot from a similar angle, features around the vehicle roof and front glass are the most common.

4. Visualization tool

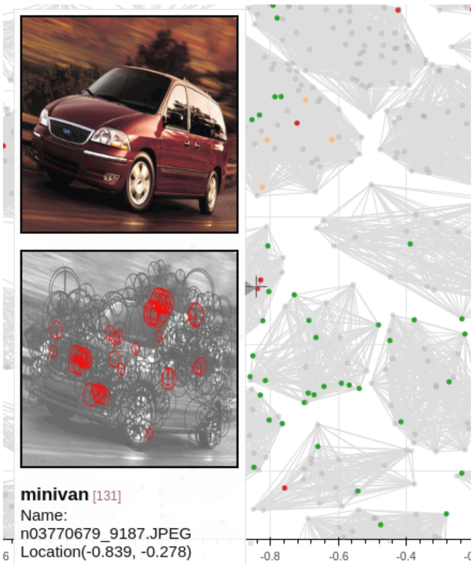
Using the visualization framework Bokeh [2], we developed an interactive tool to visually inspect synsets. It opens a pre-processed synset, showing the spatial embedding of its visual feature space using UMAP.

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

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

5. Comparing image regions

If training the visual model on full images, features in the foreground and background are treated equally. For the use-case of evaluating semantics across different concepts, this might actually be benefitial as the background includes extra semantic information, not otherwise avail-



minivan [131]
Name:
n03770679_9187.JPEG
Location(-0.839, -0.278)

Fig. 4 Visualization tool allows browsing the dataset. On mouseover, it highlights information and example images of each datapoint, including a visualization of the most common visual words in relation to neighboring images.

able in the visual data. For example, the visualization can showcase clusters of concepts, where a background plays a more important role, than the foreground. In a visual concept containing all *vehicles*, images of *helicopters* and *airplanes* might be clustered together, as the visual characteristics of clouds are visually more important, than characteristics along their chassis.

In an object recognition context, this will inevitably lead to unfavorable results, needing for image segmentation. When assessing the psycholinguistics, however, concepts might create a similar mental image, if they appear in similar situations. Therefore, an analysis of common backgrounds might help estimating properties like familiarity, concreteness, imagability, and meaningfulness.

6. Conclusion

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

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

In future work, we plan to use the results to estimate psycholinguistic ratings like familiarity, concreteness, imagability, and meaningfulness [11]. Furthermore, we want to look into visualizing other visual features than BoVW models, as well as looking into other visualization methods like heatmaps.

Acknowledgments

Parts of this research were supported by the MEXT, Grant-in-Aid for Scientific Research, and a joint research project with NII, Japan.

References

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

[2] Bokeh Development Team: *Bokeh: Python library for interactive visualization* (2014).

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

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

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

[6] Hentschel, C. and Sack, H.: What image classifiers really see — Visualizing bag-of-visual words models, *MultiMedia Modeling* (He, X., Luo, S., Tao, D., Xu, C., Yang, J. and Hasan, M. A., eds.), Springer, pp. 95–104 (2015).

[7] McInnes, L. and Healy, J.: UMAP: Uniform Manifold Approximation and Projection for dimension reduction, *ArXiv e-prints 1802.03426* (2018).

[8] Miller, G. A.: WordNet: A lexical database for English, *Comm. ACM*, Vol. 38, No. 11, pp. 39–41 (1995).

[9] Paivio, A., Yuille, J. C. and Madigan, S. A.: Concreteness, imagery, and meaningfulness values for 925 nouns, *J Exp Psychol*, Vol. 76, No. 1, pp. 1–25 (1968).

[10] Richardson, J. T. E.: Imageability and concreteness, *Bull Psychon Soc*, Vol. 7, No. 5, pp. 429–431 (1976).

[11] Wilson, M., Wilson, M., Qx, O. O. and Quinlan, P.: MRC Psycholinguistic Database: Machine Usable Dictionary, Version 2.00. (1987).

[12] Yue, H., Chen, W., Wu, X. and Wang, J.: Visualizing bag-of-words for high-resolution remote sensing image classification, *J Appl Remote Sens*, Vol. 10, No. 1 (2016).

[13] カストナーマークアウレル, 井手一郎, 川西康友, 平山高嗣, 出口大輔 and 村瀬洋: Web画像の分布に基づく単語概念の視覚的な多様性の推定, *IPSJ SIG Technical Report*, Vol. 2018-CVIM-211, No. 4 (2018).