CVIM-5

On the quantification of the mental image of visual concepts for multi-modal applications

Marc A. Kastner (NII), I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, H. Murase (Nagoya Univ.)

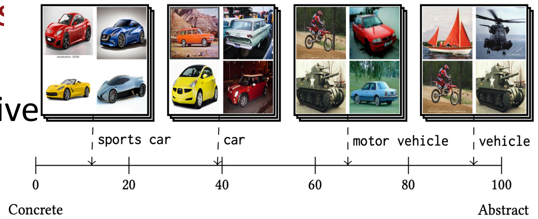
mkastner@nii.ac.jp

https://www.marc-kastner.com/

Background

- **Semantic gap problems**
  - Missing information between computer representation and human perception
  - Often an issue in word choice problems and resulting in unnatural results

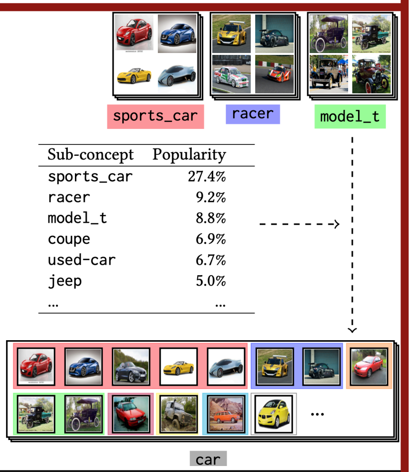
- **Psycholinguistics looks at perception of words**
  - Up to nine different measures per word …
  - … but dataset creation is manual and labor intensive



In my doctoral studies I analyzed the mental image of concepts for use in multimedia modeling

Relative measurements

- **Idea: Data mine visual features to quantify feature variety across related words**
  - E.g. compare variety of *car* vs. *sports car* vs. *vehicle*
  - Analyses quickly showed bias in existing datasets
- **Proposed method: Improve dataset by recomposing existing datasets**
  - Create hypernym datasets based on their hyponyms
  - Use a Web-based ratio to determine composition
  - Lastly, cluster feature space to determine number of visually distinct concepts

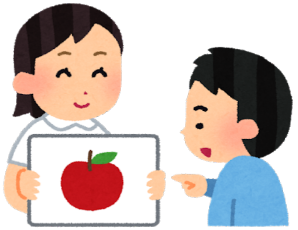


Sub-concept	Popularity
sports_car	27.4%
racer	9.2%
model_t	8.8%
coupe	6.9%
used-car	6.7%
jeep	5.0%
…	…

Corpus	Correlation (1 = best)	MSE (0 = best)
Plain ImageNet (Baseline)	0.25	10.54
Equal weighting (Comparative)	0.62	9.23
Popularity weighting (Proposed)	**0.73**	**9.01**

Core ideas

- **Try to quantify semantic gap before solving it**
  - Use visual data mining to estimate visual variety differences across datasets
  - Estimate perception of concepts without manual labor needed
- **Applications**
  - Word choice problems like retrieval or tagging
  - Increase vocabulary of psycholinguistics dictionaries



Absolute measurements

- **Idea: Data mine visual features for words in an imageability dictionary**
  - Score words from 1 (unimageable) to 7 (imageable)
  - Regress such a scoring using images
- **Proposed method: Gain visual information across low- and high-level features**
  - **Low**: Patterns or colors
  - **High**: Objects or concepts

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 (YOLO-based)	0.43	12.55
H3: Image composition (YOLO-based)	0.25	13.98
Combined (Proposed method)	**0.63**	**10.14**
Local visual variety approach [3]	-0.01	67.31