# 視覚情報と言語情報の間のセマンティックギャップを埋めるための挑戦

井手 一郎†

†名古屋大学 大学院情報学研究科 〒464-8601 愛知県名古屋市千種区不老町
E-mail: †ide@i.nagoya-u.ac.jp

## あらまし

講演者は，過去 25 年間，マルチメディアコンテンツの解析，特に視覚と言語（Vision and Language ; V&L）情報間の関係解析に取り組んできた．近年，両者に対する処理ツールの進歩や実世界データの増加により，V&L 情報間の一般的な関係を分析することが現実的になりつつある．本講演では，そのような試みのうち，1）単語や文の「心像性（Imageability）」（心理言語学の概念）を定量化し，心像性が異なるキャプションを生成する**制御可能な画像キャプショニング**と，2）人間の動きを擬態語で表現，また逆に擬態語から人間の動きを生成することができるような，**人間の動きと擬態語の関係のモデル化**について紹介する．このような研究により，V&L 情報の間によこたわる，いわゆる「セマンティックギャップ（Semantic gap）」を埋める端緒になると考えている．

## 1. 制御可能な画像キャプショニング

本研究では，心理言語学の概念である「心像性（Imageability）」に着目し，単語概念の想像しやすさ[1]を定量化することを目指している．従来，心理言語学分野では，大勢の被験者を対象とした心理実験により定量化していたが，辞書作成に多大なコストを要する．そこで，本講演では，多数のウェブ画像から抽出した画像特徴の傾向と，単語自身の特徴及びその発音など複数のモダリティから得られる情報に基づいて，自動的に定量化する手法を紹介する．

次に，このようにして任意の単語について定量化できるようになった心像性の度合と長さをパラメータとして，出力文を制御できる画像キャプショニング手法を紹介する．このような手法により，用途に応じた画像キャプションを出力できるようになる．

本研究の詳細については，文献[1–5]を参照されたい．

## 2. 人間の動きと擬態語の関係のモデル化

本講演では擬態語，特に日本語のオノマトペ（Onomatopoeia）に着目し，その発音と人間の動きとの関係をモデル化する手法を紹介する．このようなモデル化により，人間の動きを直感的に分かりやすく言語で表現できるようになるほか，直感的な表現で人間の動きを生成できるようになる．

本研究の詳細については，文献[7–9]を参照されたい．

## 謝辞

本研究の一部は，科研費（16H02846，22H03612），国立情報学研究所及びアムステルダム大学との共同研究による．

## 文　献

[1] A Paivio, JC Yuille, and SA Madigan: “Concreteness, imagery, and meaningfulness values for 925 nouns”, *J Exp Psycho*, **76**(1, Pt 2):1–25, Jan 1968.

[2] MA Kastner, K Umemura, I Ide, Y Kawanishi, T Hirayama, K Doman, D Deguchi, H Murase, and S Satoh: “Imageability- and length-controllable image captioning”, *IEEE Access*, **9**: 162951–162961, Nov 2021.

[3] MA Kastner, C Matsuhira, I Ide, and S Satoh: “A multi-modal dataset for analyzing the imageability of concepts across modalities”, In *Proc 4th IEEE Int Conf Multimed Inf Process Retr*, 213–218, Sept 2021.

[4] K Umemura, MA Kastner, I Ide, Y Kawanishi, T Hirayama, K Doman, D Deguchi, and H Murase: “Tell as you imagine: Sentence imageability-aware image captioning”, In *Proc 27th Int Conf Multimed Model*, **2**: 62–73, June 2021.

[5] C Matsuhira, MA Kastner, I Ide, Y Kawanishi, T Hirayama, K Doman, D Deguchi, and H Murase: “Imageability estimation using visual and language features”, In *Proc ACM Int Conf Multimed Retr 2020*, 306–310, Oct 2020.

[6] MA Kastner, I Ide, F Nack, Y Kawanishi, T Hirayama, D Deguchi, and H Murase: “Estimating the imageability of words by mining visual characteristics from crawled image data”, *Multimed Tools Appl*, **79**(25): 18167–18199, July 2020.

[7] H Kato, T Hirayama, K Doman, I Ide, Y Kawanishi, T Komamizu, D Deguchi, and H Murase: “Intuitive gait modeling using mimetic-words for gait description and generation”, In *Proc 5th IEEE Int Conf Multimed Inf Process Retr*, 240–245, Aug 2022.

[8] H Kato, T Hirayama, I Ide, K Doman, Y Kawanishi, D Deguchi, and H Murase: “More-natural mimetic words generation for fine-grained gait description”, In *Proc 26th Int Conf Multimed Model*, **2**: 214–225, Jan 2020.

[9] H Kato, T Hirayama, Y Kawanishi, K Doman, I Ide, D Deguchi, and H Murase: “Toward describing human gaits by onomatopoeias”, In *Proc 7th IEEE Int Workshop Anal Model Faces Gestures*, 1573–1580, Oct 2017.