Visual Features for Linguists: Basic image analysis techniques for multimodally-curious NLPers

Elia Bruni
University of Trento
elia.bruni@unitn.it

Marco Baroni
University of Trento
marco.baroni@unitn.it

Description

Features automatically extracted from images constitute a new and rich source of semantic knowledge that can complement information extracted from text. The convergence between vision- and text-based information can be exploited in scenarios where the two modalities must be combined to solve a target task (e.g., generating verbal descriptions of images, or finding the right images to illustrate a story). However, the potential applications for integrated visual features go beyond mixed-media scenarios: Because of their complementary nature with respect to language, visual features might provide perceptually grounded semantic information that can be exploited in purely linguistic domains.

The tutorial will first introduce basic techniques to encode image contents in terms of low-level features, such as the widely adopted SIFT descriptors. We will then show how these low-level descriptors are used to induce more abstract features, focusing on the well-established bags-of-visual-words method to represent images, but also briefly introducing more recent developments, that include capturing spatial information with pyramid representations, soft visual word clustering via Fisher encoding and attribute-based image representation. Next, we will discuss some example applications, and we will conclude with a brief practical illustration of visual feature extraction using a software package we developed.

The tutorial is addressed to computational linguists without any background in computer vision. It provides enough background material to understand the vision-and-language literature and the less technical articles on image analysis. After the tutorial, the participants should also be able to autonomously incorporate visual features in their NLP pipelines using off-the-shelf tools.

Outline

1. Why image analysis?
   - The grounding problem
   - Multimodal datasets (Pascal, SUN, ImageNet and ESP-Game)
2. Extraction of low-level features from images
   - Challenges (viewpoint, illumination, scale, occlusion, etc.)
   - Feature detectors
   - Feature descriptors
3. Visual words for higher-level representation of visual information
   - Constructing a vocabulary of visual words
   - Classic Bags-of-visual-words representation
   - Recent advances
   - Computer vision applications: Object recognition and emotion analysis
4. Going multimodal: Example applications of visual features in NLP
   - Generating image descriptions
   - Semantic relatedness
   - Modeling selectional preference

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, page 1,
Sofia, Bulgaria, August 4-9 2013.