Abstract

Automatically extracting structured information from images is becoming increasingly important as the amount of available visual data grows. We present an approach to spatial relation prediction in images which makes use of two kinds of object properties: spatial characteristics and lexical knowledge extracted from corpora and WordNet. These properties are formalised as predicates in first-order semantic models, allowing for integrated reasoning. Our focus is on the prediction of three spatial relations: `part of`, `touching`, and `supports`. We frame the prediction as a supervised classification task and obtain our gold standard labels via crowdsourcing. Results show that a combination of spatial and lexical knowledge performs better than using spatial and lexical information in isolation. While spatial information is important throughout, relations differ in their preferences for lexical knowledge (for instance, `part of` relies heavily on part meronymy information, while `supports` benefits from preposition statistics derived from a large corpus). We conclude that knowing *what* objects are (lexical knowledge) can improve prediction of spatial relations compared to only knowing *where* they are.