Challenges for annotating images for sense disambiguation

Cecilia Ovesdotter Alm
Dept. of Linguistics
University of Illinois, UC
ebbaalm@uiuc.edu

Nicolas Loeff
Dept. of Computer Science
University of Illinois, UC
loeff@uiuc.edu

David A. Forsyth
Dept. of Computer Science
University of Illinois, UC
daf@uiuc.edu

Abstract

We describe an unusual data set of thousands of annotated images with interesting sense phenomena. Natural language image sense annotation involves increased semantic complexities compared to disambiguating word senses when annotating text. These issues are discussed and illustrated, including the distinction between word senses and iconographic senses.

1 Introduction

We describe a set of annotated images, each associated with a sense of a small set of words. Building this data set exposes important sense phenomena which not only involve natural language but also vision. The context of our work is *Image Sense Discrimination* (ISD), where the task is to assign one of several senses to a web image retrieved by an ambiguous keyword. A companion paper introduces the task, presents an unsupervised ISD model, drawing on web page text and image features, and shows experimental results (Loeff et al., 2006). The data was subject to single-annotator labeling, with verification judgements on a part of the data set as a step toward studying agreement. Besides a test bed for ISD, the data set may be applicable to e.g. multimodal word sense disambiguation and cross-language image retrieval. The issues discussed concern concepts, and involve insights into semantics, perception, and knowledge representation, while opening up a bridge for interdisciplinary work involving vision and NLP.

2 Related work

The complex relationship between annotations and images has been explored by the library community, who study management practices for image collections, and by the computer vision community, who would like to provide automated image retrieval tools and possibly learn object recognition methods.

Commercial picture collections are typically annotated by hand, e.g. (Enser, 1993; Armitage and Enser, 1997; Enser, 2000). Subtle phenomena can make this very difficult, and content vs. interpretation may differ; an image of the Eiffel tower could be annotated with *Paris* or even *love*, e.g. (Armitage and Enser, 1997), and the resulting annotations are hard to use, cf. (Markkula and Sormunen, 2000), or Enser's result that a specialized indexing language gives only a "blunt pointer to regions of the Hulton collections", (Enser, 1993), p. 35.

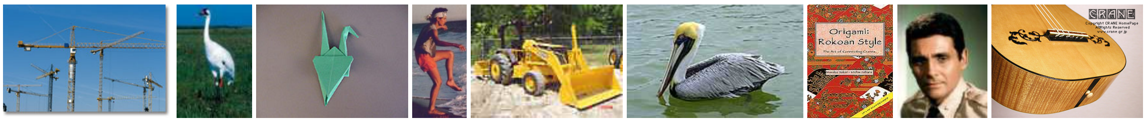
Users of image collections have been well studied. Important points for our purposes are: Users request images both by object kinds, and individual identities; users request images both by what they depict and by what they are about; and that text associated with images is extremely useful in practice, newspaper archivists indexing largely on captions (Markkula and Sormunen, 2000).

The computer vision community has studied methods to predict annotations from images, e.g. (Barnard et al., 2003; Jeon et al., 2003; Blei and Jordan, 2002). The annotations that are predicted most successfully tend to deal with materials whose identity can be determined without shape analysis, like *sky*, *sea* and the like. More complex annotations remain difficult. There is no current theory of word sense in this context, because in most current collections, words appear in the most common sense only. Sense is known to be important, and image information can disambiguate word senses (Barnard and Johnson, 2005).

Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006, pages 1–4,
Sydney, July 2006.

Word (#Annot. images)	QueryTerms	Senses	Coverage	Examples of visual annotation cues
BASS (2881)	5: bass, bass guitar, bass instrument, bass fishing, sea bass	1. **fish**	35%	any fish, people holding catch
		2. **musical instrument**	28%	any bass-looking instrument, playing
		3. related: fish	10%	fishing (gear, boats, farms), rel. food, rel. charts/maps
		4. related: musical instrument	8%	speakers, accessories, works, chords, rel. music
		5. unrelated	12%	miscellaneous (above senses not applicable)
		6. people	7%	faces, crowds (above senses not applicable)
CRANE (2650)	5: crane, construction cranes, whooping crane, sandhill crane, origami cranes	1. **machine**	21%	machine crane, incl. panoramas
		2. **bird**	26%	crane bird or chick
		3. **origami**	4%	origami bird
		4. related: machine	11%	other machinery, construction, motor, steering, seat
		5. related: bird	11%	egg, other birds, wildlife, insects, hunting, rel. maps/charts
		6. related: origami	1%	origami shapes (stars, pigs), paper folding
		7. people	7%	faces, crowds (above senses not applicable)
		8. unrelated	18%	miscellaneous (above senses not applicable)
		9. **karate**	1%	martial arts
SQUASH (1948)	10: squash+: rules, butternut, vegetable, grow, game of, spaghetti, winter, types of, summer	1. **vegetable**	24%	squash vegetable
		2. **sport**	13%	people playing, court, equipment
		3. related:vegetable	31%	agriculture, food, plant, flower, insect, vegetables
		4. related:sport	6%	other sports, sports complex
		5. people	10%	faces, crowds (above senses not applicable)
		6. unrelated	16%	miscellaneous (above senses not applicable)

Table 1: Overview of annotated images for three ambiguous query terms, inspired by the WSD literature. For each term, the number of annotated images, the expanded query retrieval terms (taken terms from `askjeeves.com`), the senses, their distribution coverage, and rough sample annotation guidelines are provided, with core senses marked in bold.



(a) *machine* (b) *bird* (c) *origami* (d) *karate* (e) *rel. to a* (f) *rel. to b* (g) *rel. to c* (h) *people* (i) *unrel.*

Figure 1: CRANE images with clear senses: (a-d) *core* senses, (e-g) *related* senses, (h) *people* and (i) *unrelated*. Related senses are associated with the semantic field of a core sense, but the core sense is visually absent or undeterminable.

3 Data set

The data set has images retrieved from a web search engine. We deliberately focused on three keywords, which cover a range of phenomena in semantic ambiguity: BASS, CRANE, and SQUASH. Table 1 gives an overview of the data set, annotated by one author (CA).[1] The webpage was not considered to avoid bias, given the ISD task.

For each query, 2 to 4 core word senses were distinguished from inspecting the data using common sense. We chose this approach rather than ontology senses which tend to be incomplete or too specific for our purposes. For example, the *origami* sense of CRANE is not included in WordNet under CRANE, but for BASS three different senses appear with fish. WordNet contains *bird* as part of the description for the separate entry *origami*, and some query expansion terms are hyponyms which occur as separate WordNet entries (e.g. *bass guitar*, *sea bass*, *summer squash*). Images may show multiple objects; a general strategy preferred a core sense if it was included.

An additional complication is that given that the images are retrieved by a search engine there is no guarantee that they depict the query term, so additional senses were introduced. Thus, for most core senses, a RELATED label was included for meanings related to the semantic field of a core sense. Also, a PEOPLE label was included since such images may occur due to how people take pictures (e.g. portraits of persons, group pictures, or other representations of people outside core and related senses). An UNRELATED label accounted for images that did not fit other labels, or were irrelevant or undeterminable. In fact, distinguishing between PEOPLE and UNRELATED was not always straightforward. Fig. 1 shows examples of CRANE when sense assignment was quite straightforward. However, distinguishing image senses was often not this clear. In fact, many border-line cases occurred when one could argue for different label assignments. Also, annotation cues are subject to interpretation, and disagreements between judges are expected. They simply reflect that image senses are located on a semantic continuum.

4 Why annotating image senses is hard

In general, annotating images involves special challenges, such as what to annotate and how extensively. We assign an image one sense. Nevertheless, compared to disambiguating a word, several issues are added for annotation. As noted above, a core sense may not occur, and judgements are characterized by increased subjectivity, with semantics beyond prototypical and peripheral

[1] We call the data set the *UIUC-ISD data set*. It is currently at `http://www.visionpc.cs.uiuc.edu/isd/`.

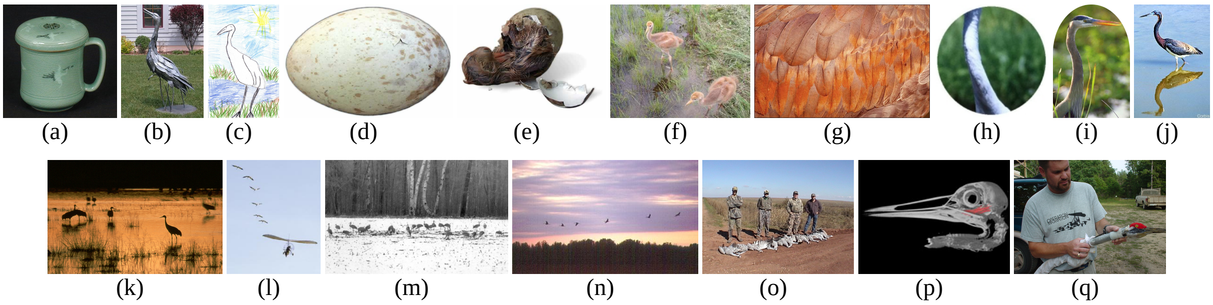


Figure 2: Annotating images is often challenging for different reasons. Are these images of CRANE birds? (a-c) depiction (d-f) gradient change (g-h) partial display (i-j) domain knowledge (k) unusual appearance (l-n) distance (o-q) not animate.

exemplars. Also, the disambiguating context is limited to image contents, rather than collocations of an ambiguous token. Fig. 2 illustrates selected challenging judgement calls for assigning or not the bird sense of CRANE, as discussed below.

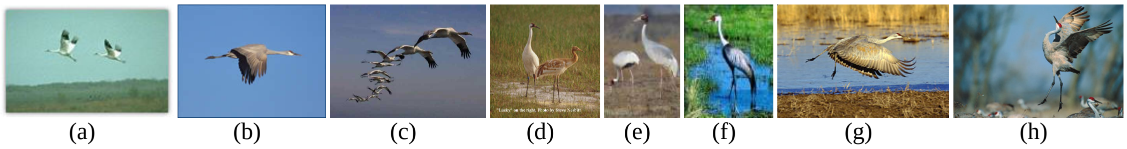
Depiction: Images may include man-made depictions of an object in artistic depictions, and the question is whether this counts as the object or not, e.g. Fig. 2(a-c). **Gradient changes:** Recognition is complicated by objects taking different forms and shapes, cf. the insight by (Labov, 1973) on gradual categories.[2] For example, as seen in Fig. 2(d-f), birds change with age; an egg may be a bird, but a chick is, as is a fledgeling. **Partial display:** Objects may be rendered in incomplete condition. For example, Fig. 2(g-h) show merely feathers or a bird neck. **Domain knowledge:** People may disagree due to differences in domain knowledge, e.g. some non-experts may have a difficult time determining whether or not other similar bird species can be distinguished from a bird crane, cf. Fig. 2(i-j). This also affected annotations' granularity depending on keyword, see Table 1's example cues. **Unusual appearance:** Objects may occur in less frequent visual appearance, or lack distinguishing properties. For instance, Fig. 2(k) illustrates how sunset background masks birds' color information. **Scale:** The distance to objects may render them unclear and influence judgement accuracy, and people may differ in the degree of certainty required for assigning a sense. For example, Fig. 2(l-n) show flying or standing potential cranes at distance. **Animate:** Fig. 2(o-q) raise the question whether dead, skeletal, or artificial objects are instantiations or not. Other factors complicating the annotation task include image **crowdedness** disguising objects, certain entities having less **salience**, and lacking or unclear **reference to object proportions**. Senses may also be **etymologically** related or **blend** occasionally, or be guided by **cultural** interpretations, and so on.

Moreover, related senses are meant to capture images associated with the semantic field of a core sense. However, because the notion and borders of a semantic field are non-specific, **related senses are tricky**. Annotators may build associations quite wildly, based on personal experience and opinion, thus what is or is not a related sense may very quickly get out of hand. For instance, a person may by association reason that if bird cranes occur frequently in fields, then an image of a field alone should be marked as related. To avoid this, guidelines attempted to restrict related senses, as exemplified in Table 1, with some data-driven revisions during the annotation process. However, guidelines are also based on judgement calls. Besides, for abstract concepts like LOVE, differentiating core versus related sense is not really valid.

Lastly, an additional complexity of image senses is that in addition to traditional word senses, images may also capture repeatedly occurring **iconographic** patterns or senses. As illustrated in Fig. 3, the iconography of flying cranes is quite different from that of standing cranes, as regards motion, shape, identity, and color of figure and ground, respectively. Mixed cases also occur, e.g. when bird cranes are taking off or are about to land in relation to flight. Iconographic senses may compare to more complex linguistic structures than nominal categories, e.g. a modified NP or clause, but are represented by image properties.

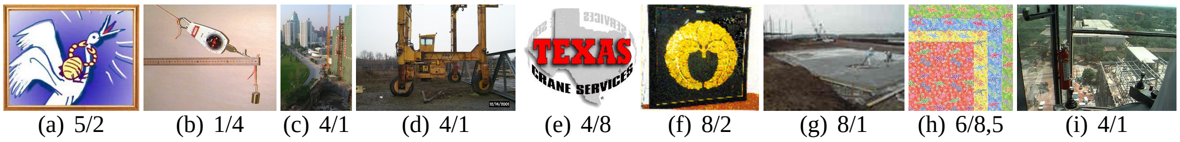
A policy for annotating iconographic senses is still lacking. Image groups based on iconographic senses seem to provide increased visual and semantic harmony for the eye, but experiments are needed to confirm how iconographic senses correspond to humans' perception of semantic image similarity, and at what level of semantic differen-

[2] Function or properties may also influence (Labov, 1973).



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3: Iconographic bird CRANE senses: (a-c) *flying cranes*, (d-f) *standing cranes*, and (g-h) *mixed cases* in-between.



(a) 5/2 (b) 1/4 (c) 4/1 (d) 4/1 (e) 4/8 (f) 8/2 (g) 8/1 (h) 6/8,5 (i) 4/1

Figure 4: Disagreement examples (sense numbers in Table 1): (a) crane or other bird? (b) toy crane or scales? (c) crane or other steel structure/elevator? (d) crane or other machine? (e) company is related or not? (f) bird or abstract art? (g) crane in background or not? (h) origami-related paper? (i) inside of crane? (and is inside sufficient to denote image as machine crane?)

tiation they become relevant for sense assessment.

Lastly, considering the challenges of image annotation, it is interesting to look at annotation disagreements. Thus, another author (NL) inspected CRANE annotations, and recorded disagreement candidates, which amounted to 5%. Rejecting or accepting a category label seems less hard than independent annotation but still can give insights into disagreement tendencies. Several disagreements involved a core category vs. its related label vs. unrelated, rather than two core senses. Also, some disagreement candidates had tiny, fuzzy, partial or peripheral potential sense objects, or lacked distinguishing object features, so interpretation became quite idiosyncratic. The disagreement candidates were discussed together, resulting in 2% being true disagreements, 2% false disagreements (resolved by consensus on CA's labels), and 1% annotation mistakes. Examples of true disagreements are in Fig. 4. Often, both parties could see each others' points, but opted for another interpretation; this confirms that border lines tend to merge, indicating that consistency is challenging and not always guaranteed. As the annotation procedure advances, criteria may evolve and modify the fuzzy sense boundaries.

5 Conclusion

This work draws attention to the need for considering natural language semantics in multi-modal settings. Annotating image senses adds increased complexity compared to word-sense annotation in text due to factors such as image properties, subjective perception, and annotator domain-knowledge. Moreover, the concept of related senses as well as iconographic senses go beyond and diversify the notion of word sense. In the future, we would like to perform experimentation with human subjects to explore both similarity judgements for image pairs or groups, as well as issues in interannotator agreement for image disambiguation, and, finally, to better understand the role of iconography for semantic interpretation.

6 Acknowledgements

Thanks to anonymous reviewers, R. Girju and R. Sproat for feedback. Any fallacies are our own.

References

L. H. Armitage and P. G. B. Enser. 1997. Analysis of user need in image archives. *J. of Inform. Sci.*, 23(4):287–299.

K. Barnard and M. Johnson. 2005. Word sense disambiguation with pictures. *Artif. Intel.*, 167:13–30.

K. Barnard, P. Duygulu, N. Freitas, D. Forsyth, D. Blei, and M. I. Jordan. 2003. Matching words and pictures. *J. of Mach. Learn. Research*, 3:1107–1135.

D. M. Blei and M. I. Jordan. 2002. Modeling annotated data. Technical Report CSD-02-1202, Div. of Computer Science, Univ. of California, Berkeley.

P. G. B. Enser. 1993. Query analysis in a visual information retrieval context. *J. of Doc. and Text Management*, 1(1):25–52.

P. G. B. Enser. 2000. Visual image retrieval: seeking the alliance of concept based and content based paradigms. *J. of Inform. Sci.*, 26(4):199–210.

J. Jeon, V. Lavrenko, and R. Manmatha. 2003. Automatic image annotation and retrieval using crossmedia relevance models. In *SIGIR*, pages 119–126.

W. Labov. 1973. The boundaries of words and their meanings. In C. J. Baily and R. Shuy, editors, *New ways of analyzing variation in English*, pages 340–373. Washington D.C: Georgetown Univ. Press.

N. Loeff, C. O. Alm, and D. A. Forsyth. 2006. Discriminating image senses by clustering with multimodal features. In *ACL (forthcoming)*.

M. Markkula and E. Sormunen. 2000. End-user searching challenges indexing practices in the digital newspaper photo archive. *Inform. Retr.*, 1:259–285.