Discriminating image senses by clustering with multimodal features

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

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

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

Abstract

We discuss *Image Sense Discrimination* (ISD), and apply a method based on spectral clustering, using multimodal features from the image and text of the embedding web page. We evaluate our method on a new data set of annotated web images, retrieved with ambiguous query terms. Experiments investigate different levels of sense granularity, as well as the impact of text and image features, and global versus local text features.

1 Introduction and problem clarification

Semantics extends beyond words. We focus on *image sense discrimination* (ISD)[1] for web images retrieved from ambiguous keywords, given a multimodal feature set, including text from the document which the image was embedded in. For instance, a search for CRANE retrieves images of crane machines, crane birds, associated other machinery or animals etc., people, as well as images of irrelevant meanings. Current displays for image queries (e.g. Google or Yahoo!) simply list retrieved images in any order. An application is a user display where images are presented in semantically sensible clusters for improved image browsing. Another usage of the presented model is automatic creation of sense discriminated image data sets, and determining available image senses automatically.

ISD differs from word sense discrimination and disambiguation (WSD) by increased complexity in several respects. As an initial complication, both **word and iconographic sense distinctions** matter. Whereas a search term like CRANE can refer to, e.g. a MACHINE or a BIRD; iconographic distinctions could additionally include birds standing, vs. in a marsh land, or flying, i.e. sense-distinctions encoded by further descriptive modification in text. Therefore, as the number of text senses grow with corpus size, the iconographic senses grow even faster, and enumerating iconographic senses is extremely challenging; especially since dictionary senses do not capture iconographic distinctions. Thus, we focus on image-driven word senses for ISD, but we acknowledge the importance of iconography for visual meaning.

Also, an image often **depicts a related meaning**. E.g. a picture retrieved for SQUASH may depict a squash bug (i.e. an insect on a leaf of a squash plant) instead of a squash vegetable, whereas this does not really apply in WSD, where each instance concerns the ambiguous term itself. Therefore, it makes sense to consider the division between **core sense, related sense, and unrelated sense** in ISD, and, as an additional complication, their boundaries are often blurred. Most importantly, whereas the one-sense-per-discourse assumption (Yarowsky, 1995) also applies to discriminating images, there is **no guarantee of a local collocational or co-occurrence context** around the target image. Design or aesthetics may instead determine image placement. Thus, considering local text around the image may not be as helpful as local context is for standard WSD. In fact, the **query term may even not occur** in the text body. On the other hand, one can assume that an image spotlights the web page topic and that it highlights important document information. Also, images mostly depict concrete senses. Lastly, ISD from web data is complicated by web pages being more domain-independent than news wire, the fa-

[1] Cf. (Schütze, 1998) for a definition of sense discrimination in NLP.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 547–554,
Sydney, July 2006.

(a) *squash flower* (b) *tennis?* (c) *hook* (d) *food* (e) *bow* (f) *speaker*

Figure 1: **Example RELATED images** for (a) *vegetable* and (b) *sports* senses for SQUASH, and for (c-d) *fish* and (e-f) *musical instrument* for BASS. Related senses are associated with the semantic field of a core sense, but the core sense is visually absent or undeterminable.

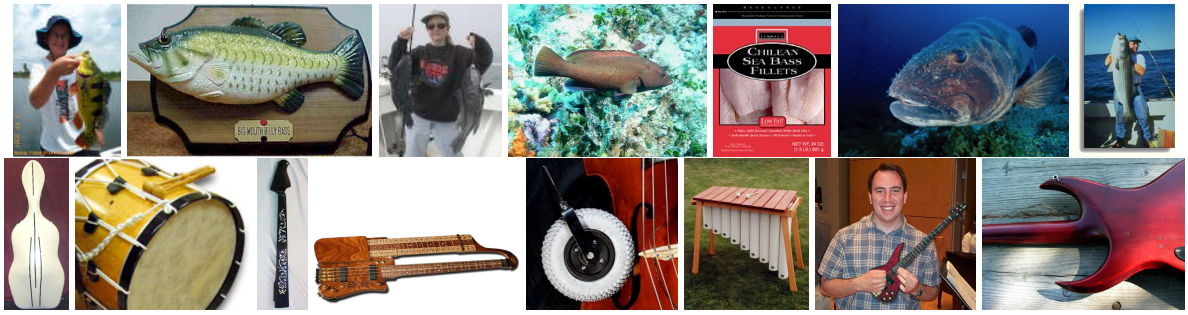


Figure 2: ***Which fish or instruments are* BASS?** Image sense annotation is more vague and subjective than in text.

vored corpus for WSD. As noted by (Yanai and Barnard, 2005), whereas current image retrieval engines include many irrelevant images, a data set of web images gives a more real-world point of departure for image recognition.

Outline Section 2 discusses the corpus data and image annotation. Section 3 presents the feature set and the clustering model. Subsequently, section 4 introduces the evaluation used, and discusses experimental work and results. In section 5, this work is positioned with respect to previous work. We conclude with an outline of plans for future work in section 6.

2 Data and annotation

Yahoo!'s image query API was used to obtain a corpus of pairs of semantically ambiguous images, in thumbnail and true size, and their corresponding web sites for three ambiguous keywords inspired by (Yarowsky, 1995): BASS, CRANE, and SQUASH. We apply query augmentation (cf. Table 1), and exact duplicates were filtered out by identical image URLs, but cases occurred where both thumbnail and true-size image were included. Also, some images shared the same webpage or came from the same site. Generally, the latter gives important information about shared discourse topic, however the images do not necessarily depict the same sense (e.g. a CRANE bird vs. a meadow), and image features can separate them into different clusters.

Annotation overview The images were annotated with one of several labels by one of the authors out of context (*without* considering the web site and its text), after applying text-based filtering (cf. section 3.1). For annotation purposes, images were numbered and displayed on a web page in thumbnail size. In case the thumbnail was not sufficient for disambiguation, the image linked at its true size to the thumbnail was inspected.[2] The true-size view depended on the size of the original picture and showed the image and its name. However, the annotator tried to resist name influence, and make judgements based just on the image. For each query, 2 to 4 core word senses (e.g. *squash vegetable* and *squash sport* for SQUASH) were distinguished from inspecting the data. However, because "context" was restricted to the image content, and there was no guarantee that the image actually depicts the query term, additional annotator senses were introduced. Thus, for most core senses, a RELATED label was included, accounting for meanings that seemed related to core meaning but lacked a core sense object in the image. Some examples for RELATED senses are in Fig. 1. In addition, for each query term, a PEOPLE label was included because such images are common due to the nature of how people take pictures (e.g. portraits of persons or group pictures of crowds, when core or related senses did not apply), as was an

[2] We noticed a few cases where Yahoo! retrieved a thumbnail image different from the true size image.

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, crowd (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, crowd (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, crowd (above senses not applicable)
		6. unrelated	16%	miscellaneous (above senses not applicable)

Table 1: **Web images for three ambiguous query terms** were annotated manually out of context (*without* considering the web page document). For each term, the number of annotated images, the query retrieval terms, the senses, their distribution, and rough sample annotation guidelines are provided, with core senses marked in bold face. Because image retrieval engines restrict hits to 1000 images, query expansion was conducted by adding narrowing query terms from `askjeeves.com` to increase corpus size. We selected terms relevant to core senses, i.e. the main discrimination phenomenon.

UNRELATED label for irrelevant images which did not fit other labels or were undeterminable.

For a human annotator, even when using more natural word senses, assigning sense labels to images based on image alone is more challenging and subjective than labeling word senses in textual context. First of all, the annotation is heavily dependent on **domain-knowledge** and it is not feasible for a layperson to recognize fine-grained semantics. For example, it is straightforward for the layperson to distinguish between a robin and a crane, but determining whether a given fish should have the common name *bass* applied to it, or whether an instrument is indeed a bass instrument or not, is extremely difficult (see Fig. 2; e.g. deciding if a picture of a fish fillet is a picture of a fish is tricky). Furthermore, most images **display objects only partially**; for example just the neck of a classical double bass instead of the whole instrument. In addition, **scaling, proportions, and components** are key cues for object discrimination in real-life, e.g. for singling out an electric bass from an electric guitar, but an image may not provide these detail. Thus, **senses are even fuzzier** for ISD than WSD labeling. Given that laypeople are in the majority, it is fair to assume their perspective and naiveness. This latter fact also led to annotations' level of specificity differing according to search term. Annotation criteria depended on the keyword term and its senses and their coverage, as shown in Table 1. Nevertheless, several border-line cases for label assignment occurred. Considering that the annotation task is quite subjective, this is to be expected. In fact, one person's labeling often appears as justifiable as a contradicting label provided by another person. We explore the vagueness and subjective nature of image annotation further in a companion paper (Alm, Loeff, Forsyth, 2006).

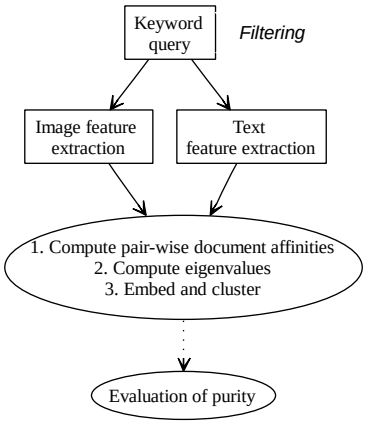


Figure 3: Overview of algorithm

3 Model

Our goal is to provide a mapping between images and a set of iconographically coherent clusters for a given query word, in an unsupervised framework. Our approach involves extracting and weighting unordered bags-of-words (BOWs; henceforth) features from the webpage text, simple local and global features from the image, and running spectral clustering on top. Fig. 3 shows an overview of the implementation.

3.1 Feature extraction

Document and text filtering A pruning process was used to filter out image-document pairs based on e.g. language specification, exclusion of *"Index of"* pages, pages lacking an extractable target image, or a cutoff threshold of number of tokens in the body. For remaining documents, text was preprocessed (e.g. lower-casing, removing punctuation, tokens being very short, having numbers or no vowels, etc.). We used a stop word list, but avoided stemming to make the algorithm language independent in other respects. When using image features, grayscale images (no color histograms) and images without salient regions (no keypoints detected) were also removed.

Text features We used the following BOWs: (a) tokens in the page *body*; (b) tokens in a ± 10 window around the target image (if multiple, the first was considered); (c) tokens in a ± 10 window around any instances of the query keyword (e.g. *squash*); (d) tokens of the target image's *alt* attribute; (e) tokens of the *title* tag; (f) some *meta* tokens.[3] Tf-idf was applied to a weighted average of the BOWs. Webpage design is flexible, and some inconsistencies and a certain degree of noise remained in the text features.

Image features Given the large variability in the retrieved image set for a given query, it is difficult to model images in an unsupervised fashion. Simple features have been shown to provide performance rivaling that of more elaborate models in object recognition (Csurka et al, 2004) and (Chapelle, Haffner, and Vapnik, 1999), and the following image bags of features were considered:

Bags of keypoints: In order to obtain a compact representation of the textures of an image, patches are extracted automatically around interesting regions or *keypoints* in each image. The keypoint detection algorithm (Kadir and Brady, 2001) uses a saliency measure based on entropy to select regions. After extraction, keypoints were represented by a histogram of gradient magnitude of the pixel values in the region (SIFT) (Lowe, 2004). These descriptors were clustered using a Gaussian Mixture with ≈ 300 components, and the resulting global patch *codebook* (i.e. histogram of codebook entries) was used as lookup table to assign each keypoint to a codebook entry.

[3] Adding to META *content*, *keywords* was an attribute, but is irregular. Embedded BODY pairs are rare; thus not used.

Color histograms: Due to its similarity to how humans perceive color, HSV (hue, saturation, brightness) color space was used to bin pixel color values for each image. Eight bins were used per channel, obtaining an 8^3 dimensional vector.

3.2 Measuring similarity between images

For the BOWs text representation, we use the common measure of *cosine similarity* (cs) of two *tf-idf* vectors (Jurafsky and Martin, 2000). The cosine similarity measure is also appropriate for keypoint representation as it is also an unordered bag. There are several measures for histogram comparison (i.e. $L1$, χ^2). As in (Fowlkes et al, 2004) we use the χ^2 distance measure between histograms h_i and h_j.

$$\chi^2{}_{i,j} = \frac{1}{2}\sum_{k=1}^{512}\frac{(h_i(k)-h_j(k))^2}{h_i(k)+h_j(k)} \quad (1)$$

3.3 Spectral Clustering

Spectral clustering is a powerful way to separate non-convex groups of data. Spectral methods for clustering are a family of algorithms that work by first constructing a pairwise-affinity matrix from the data, computing an eigendecomposition of the data, embedding the data into this low-dimensional manifold, and finally applying traditional clustering techniques (i.e. k-means) to it.

Consider a graph with a set of n vertices each one representing an image document, and the edges of the graph represent the pairwise affinities between the vertices. Let W be an $n \times n$ symmetric matrix of pairwise affinities. We define these as the Gaussian-weighted distance

$$W_{ij} = \exp\left(-\alpha^t(1-cs^t_{i,j}) - \alpha^k(1-cs^k_{i,j}) - \alpha^c\chi^2_{i,j}\right), \quad (2)$$

where $\{\alpha^t, \alpha^k, \alpha^c\}$ are scaling parameters for text, keypoints, and color features.

It has been shown that the use of multiple eigenvectors of W is a valid space onto which the data can be embedded (Ng, Jordan, Weiss, 2002). In this space *noise* is reduced while the most significant affinities are preserved. After this, any traditional clustering algorithm can be applied in this new space to get the final clusters. Note that this is a nonlinear mapping of the *original* space. In particular, we employ a variant of k-means, which includes a *selective* step that is quasi-optimal in a Vector Quantization sense (Ueda and Nakano, 1994). It has the added advantage of being more

robust to initialization than traditional k-means. The algorithm follows,

1. For given documents, compute the affinity matrix W as defined in equation 2.
2. Let D be a diagonal matrix whose (i, i)-th element is the sum of W's i-th row, and define $\mathcal{L} = D^{-1/2}WD^{-1/2}$.
3. Find the k largest eigenvectors V of $\mathcal{L}$.
4. Define E as V, with normalized rows.
5. Perform clustering on the columns of E, which represent the embedding of each image into the new space, using a *selective* step as in (Ueda and Nakano, 1994).

Why Spectral Clustering? Why apply a variant of k-means in the embedded space as opposed to the *original* feature space? The k-means algorithm cannot separate non-convex clusters. Furthermore, it is unable to cope with noisy dimensions (this is especially true in the case of the text data) and highly non-ellipsoid clusters. (Ng, Jordan, Weiss, 2002) stated that spectral clustering outperforms k-means not only on these high dimensional problems, but also in low-dimensional, multi-class data sets. Moreover, there are problems where Euclidean measures of distance required by k-means are not appropriate (for instance histograms), or others where there is not even a natural vector space representation. Also, spectral clustering provides a simple way of combining dissimilar vector spaces, like in this case text, keypoint and color features.

4 Experiments and results

In the first set of experiments, we used all features for clustering. We considered three levels of sense granularity: (1) all senses (*All*), (2) merging related senses with their corresponding core sense (*Meta*), (3) just the core senses (*Core*). For experiments (1) and (2), we used 40 clusters and all labeled images. For (3), we considered only images labeled with core senses, and thus reduced the number of clusters to 20 for a more fair comparison. Results were evaluated according to global cluster purity, cf. Equation 3.[4]

$$\text{Global purity} = \sum_{\text{clusters}} \frac{\text{\# of most common sense in cluster}}{\text{total \# images}} \quad (3)$$

[4]Purity did not include the small set of outlier images, defined as images whose ratio of distances to the second closest and closest clusters was below a threshold.

Word	All senses	Meta senses	Core senses
BASS	6 senses	4 senses	2 senses
Median	0.60	0.73	0.94
Range	0.03	0.02	0.02
Baseline	0.35	0.45	0.55
CRANE	9 senses	6 senses	4 senses
Median	0.49	0.65	0.86
Range	0.05	0.07	0.07
Baseline	0.27	0.37	0.50
SQUASH	6 senses	4 senses	2 senses
Median	0.52	0.71	0.94
Range	0.03	0.04	0.03
Baseline	0.32	0.56	0.64

Table 2: **Median and range of global clustering purity** for 5 runs with different initializations. For each keyword, the table lists the number of senses, median, and range of global cluster purity, followed by the baseline. **All** senses used the full set of sense labels and 40 clusters. **Meta** senses merged core senses with their respective related senses, considering all images and using 40 clusters. **Core** senses were clustered into 20 clusters, using only images labeled with core sense labels. Purity was stable across runs, and peaked for **Core**. The baseline reflected the frequency of the most common sense.

Word	Img	TxtWin	BodyTxt	Baseline
BASS				
Median	0.71	0.83	0.93	0.55
Range	0.05	0.03	0.05	
CRANE				
Median	0.61	0.84	0.85	0.50
Range	0.07	0.04	0.05	
SQUASH				
Median	0.71	0.91	0.96	0.64
Range	0.05	0.04	0.03	

Table 3: **Global and local features' performance**. **Core sense** images were grouped into 20 clusters, on the basis of individual feature types, and global cluster purity was measured. The table lists the median and range from 5 runs with different initializations. **Img** included just image features; **TxtWin** local tokens in a ± 10 window around the target image anchor; **BodyTxt** global tokens in the page BODY; and **Baseline** uses the most common sense. Text performed better than image features, and global text appeared better than local. All features performed above the baseline.

Median and range results are reported for five runs, given each condition, comparing against the baseline (i.e. choosing the most common sense). Table 2 shows that purity was surprisingly good, stable across query terms, and that it was highest when only core sense data was considered. In addition, purity tended to be slightly higher for BASS, which may be related to the annotator being less confident about its fine-grained sense distinctions, and thus less strict for assigning core sense labels for this query term.[5] In addition, we looked at the relative performance of individual global and local features using 20 clusters and only core

[5]A slightly modified HTML extractor yielded similar results ($\pm$0-2% median, $\pm$0-5% range cf. to Tables 2 - 4).

Figure 4: **First 30 images from a CRANE BIRD cluster** consisting of 81 images in the median run. Individual cluster purity for all senses was 0.67, and for meta senses 0.83. Not all clusters were as pure as this one; global purity for all 40 cluster was 0.49. This cluster appeared to show some iconography; mostly standing cranes. Interestingly, another cluster contained several images of flying cranes. Most weighted tokens: *cranes whooping birds wildlife species*. Table 1 has sense labels.

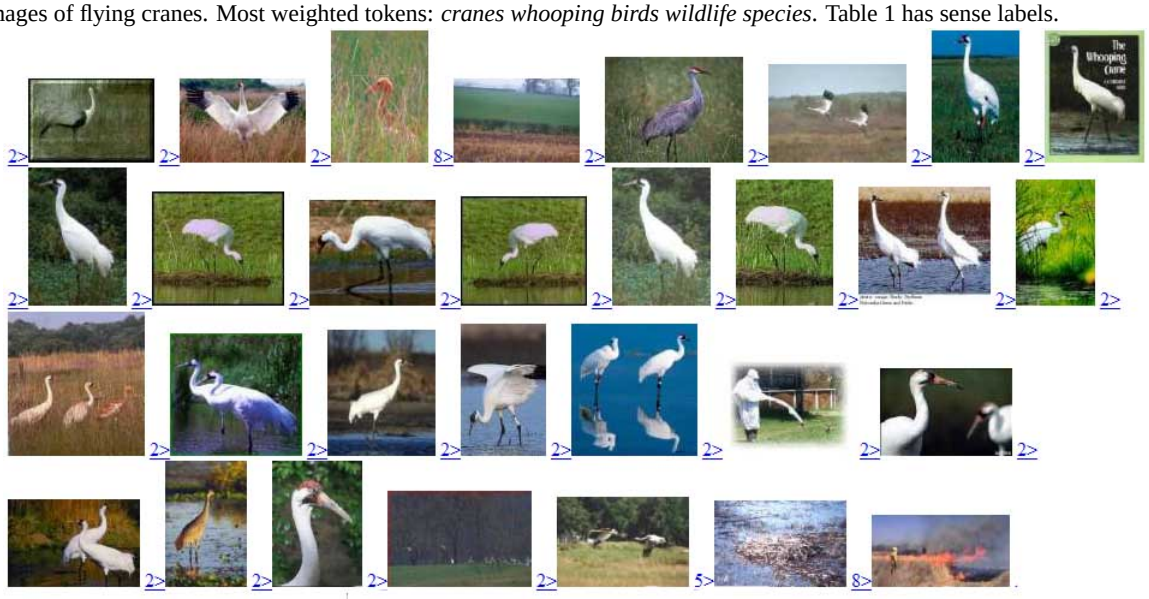
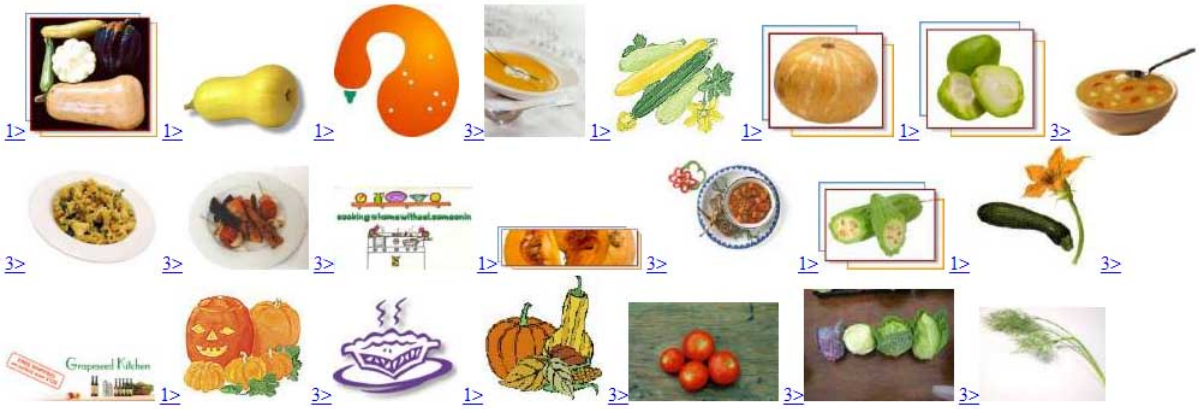


Figure 5: **Global purity does not tell the whole story** SQUASH VEGETABLE cluster of 22 images in the median run. Individual cluster purity for all senses was 0.5, and for meta senses 1.0. Global purity for all 40 cluster was 0.52. This cluster both shows visually coherent images, and a sensible meta semantic field. Most weighted tokens: *chayote calabaza add bitter cup*. Presumably, some tokens reflect the vegetable's use within the cooking domain.



sense data based on a particular feature. Table 3 shows that global text features were most informative (although not homogenously), but also that each feature type performed better than the baseline in isolation. This indicates that an optimal feature combination may improve over current performance, using manually selected parameters. In addition, purity is not the whole story. Figs. 4 and 5 show examples of two selected interesting clusters obtained for CRANE and SQUASH, respectively, using combined image and text features and all individual senses.[6] Inspection of image clusters indicated that image features, both in isolation and when used in combination, appeared to contribute to more visually balanced clusters, especially in terms of colors and shading. This shows that further exploring image features may be vital for attaining more subtle iconographic senses. Moreover, as discussed in the introduction, images are not necessarily anchored in the immediate text which they refer to. This could explain why local text features do not perform as well as global ones. Lastly, in addition, Fig. 6 shows an example of a partial cluster where the algorithm inferred a specific related sense.

We also experimented with different number of clusters for BASS. The results are in Table 4, lacking a clear trend, with comparable variation to different initializations. This is surprising, since we would expect purity to increase with number of

[6]The *UIUC-ISD data set* and results are currently at `http://www.visionpc.cs.uiuc.edu/isd/`.

Figure 6: **RELATED: SQUASH VEGETABLE cluster, consisting of 27 images.** The algorithm discovered a specific SQUASH BUG-PLANT sense, which appears iconographic. Individual cluster purity for all senses was 0.85, and individual meta purity: 1.0. Global purity for all 40 clusters: 0.52. Most weighted tokens: *bugs bug beetle leaf-footed kentucky*.

# Clusters	6	10	20	40	80
All					
Median	0.61	0.55	0.58	0.60	0.61
Range	0.03	0.05	0.03	0.03	0.04
Meta					
Median	0.75	0.70	0.70	0.73	0.72
Range	0.04	0.07	0.04	0.02	0.04

Table 4: **Impact of cluster size?** We ran BASS for different number of clusters (5 runs each with distinct initializations), and recorded median and range of global purity for all six senses of the query term, and for the four meta senses, without a clear trend.

clusters (Schütze, 1998), but may be due to the spectral clustering. Inspection showed that 6 clusters were dominated by core senses, whereas with 40 clusters a few were also dominated by RELATED senses or PEOPLE. No cluster was dominated by an UNRELATED label, which makes sense since semantic linkage should be absent between unrelated items.

5 Comparison to previous work

Space does not allow a complete review of the WSD literature. (Yarowsky, 1995) demonstrated that semi-supervised WSD could be successful. (Schütze, 1998) and (Lin and Pantel, 2002a, b) show that clustering methods are helpful in this area.

While ISD has received less attention, image categorization has been approached previously by adding text features. For example, (Frankel, Swain, and Athitsos, 1996)'s WebSeer system attempted to mutually distinguish photos, hand-drawn, and computer-drawn images, using a combination of HTML markup, web page text, and image information. (Yanai and Barnard, 2005) found that adding text features could benefit identifying relevant web images. Using text-annotated images (i.e. images annotated with relevant keywords), (Barnard and Forsyth, 2001) clustered them exploring a semantic hierarchy; similarly (Barnard, Duygulu, and Forsyth, 2002) conducted art clustering, and (Barnard and Johnson, 2005) used text-annotated images to improve WSD. The latter paper obtained best results when combining text and image features, but contrary to our findings, image features performed better in isolation than just text. They did use a larger set of image features and segmentation, however, we suspect that differences can rather be attributed to corpus type. In fact, (Yanai, Shirahatti, and Barnard, 2005) noted that human evaluators rated images obtained via a keyword retrieval method higher compared to image-based retrieval methods, which they relate to the importance of semantics for what humans regard as matching, and because pictorial semantics is hard to detect.

(Cai et al, 2004) use similar methods to rank visual search results. While their work does not focus explicitly on sense and does not provide in-depth discussion of visual sense phenomena, these do appear in, for example, figs. 7 and 9 of their paper. An interesting aspect of their work is the use of page layout segmentation to associate text with images in web documents. Unfortunately, the au-

thors only provide an illustrative query example, and no numerical evaluation, making any comparison difficult. (Wang et al, 2004) use similar features with the goal to improve image retrieval through similarity propagation, querying specific web sites. (Fuji and Ishikawa, 2005) deal with image ambiguity for establishing an online multimedia encyclopedia, but their method does not integrate image features, and appears to depend on previous encyclopedic background knowledge, limited to a domain set.

6 Conclusion

It is remarkable how high purity is, considering that we are using relatively simple image and text representation. In most corpora used to date for research on illustrated text, word sense is an entirely secondary phenomenon, whereas our data set was collected as to emphasize possible ambiguities associated with word sense. Our results suggest that a surprisingly degree of the meaning of an illustrated object is exposed on the surface.

This work is an initial attempt at addressing the ISD problem. Future work will involve learning the algorithm's parameters without supervision, and develop a semantically meaningful image taxonomy. In particular, we intend to explore the notion of iconographic senses; surprisingly good results on image classification by (Chapelle, Haffner, and Vapnik, 1999) using image features suggest that iconography plays an important role in the semantics of images. An important aspect is to enhance our understanding of the interplay between text and image features for this purpose. Also, it remains an unsolved problem how to enumerate iconographic senses, and use them in manual annotation and classification. Experimental work with humans performing similar tasks may provide increased insight into this issue, and can also be used to validate clustering performance.

7 Acknowledgements

We are grateful to Roxana Girju and Richard Sproat for helpful feedback, and to Alexander Sorokin.

References

C. O. Alm, N. Loeff, and D. Forsyth. 2006. Challenges for annotating images for sense disambiguation. *ACL workshop on Frontiers in Linguistically Annotated Corpora*.

K. Barnard and D. Forsyth. 2001. Learning the semantics of words and pictures. *ICCV*, 408–415.

K. Barnard, P. Duygulu, and D. Forsyth. 2002. Modeling the statistics of image features and associated text. *SPIE*.

K. Barnard and M. Johnson. 2005. Word sense disambiguation with pictures. *Artificial Intelligence*, 167, 13–30.

D. Cai et al. 2004. Hierarchical clustering of WWW image search results using visual, textual and link information. *ACM Multimedia*, 952-959.

O. Chapelle and P. Haffner and V. Vapnik. 1999. Support vector machines for histogram-based image classification. *IEEE Neural Networks*, 10(5), 1055–1064.

G. Csurka et al. 2004. Visual categorization with bags of keypoints. *ECCV Int. Workshop on Stat. Learning in Computer Vision*.

C. Frankel, M. Swain, and V. Athitsos. 1996. WebSeer: an image search engine for the World Wide Web. *Univ. of Chicago, Computer Science, Technical report #96-14*.

C. Fowlkes, S. Belongie, F. Chung, and J. Malik. 2004. Spectral grouping using the Nyström method. *IEEE PAMI*, 26(2),214-225.

A. Fuji and T. Ishikawa. 2005. Toward the automatic compilation of multimedia encyclopedias: associating images with term descriptions on the web. *IEEE WI*, 536-542.

D. Jurafsky and J. Martin 2000. *Speech and Language Processing*, Prentice Hall.

T. Kadir and M. Brady. 2001. Scale, saliency and image description. *Int. Journal of Computer Vision*, 45 (2):83–105.

D. Lin and P. Pantel. 2002a. Concept discovery from text. *COLING*, 577–583.

D. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2), 91–110.

A. Ng, M. Jordan, and Y. Weiss. 2002. On spectral clustering: analysis and an algorithm. *NIPS 14*.

P. Pantel and D. Lin. 2002b. Discovering word senses from text. *KDD*, 613–619.

H. Schuetze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

J. Shi and J. Malik. 2000. Normalized cuts and image segmentation. *IEEE PAMI*, 22(8):888–905.

N. Ueda. and R. Nakano. 1994. A new competitive learning approach based on an equidistortion principle for designing optimal vector quantizers. *Neural Networks*, 7(8):1211–1227.

X.-J. Wang et al. 2004. Multi-model similarity propagation and its application for image retrieval. *MM*,944–951.

K. Yanai and K. Barnard. 2005. Probabilistic web image gathering. *SIGMM*, 57–64.

K. Yanai, N. V. Shirahatti, and K. Barnard. 2005. Evaluation strategies for image understanding and retrieval. *SIGMM*, 217-226.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *ACL*, 189–196.