Using Language to Drive the Perceptual Grouping of Local Image Features

Michael Jamieson Sven Dickinson Suzanne Stevenson
University of Toronto
{jamieson, sven, suzanne}@cs.toronto.edu

Sven Wachsmuth
Bielefeld University
swachsmu@techfak.uni-bielefeld.de

Abstract

We address the problem of learning both the semantics (names) and the visual features (SIFT collections) of objects appearing in a training set of unstructured, captioned images of cluttered scenes. Prior work in applying machine translation models to learn the associations between image features and caption nouns has assumed a one-to-one correspondence between features and nouns. However, each training image may contain thousands of SIFT features belonging to multiple objects. Our challenge is two-fold: 1) grouping the SIFT features into meaningful collections, and 2) learning the object names associated with those collections. Since better collections tend to have stronger associations with object names, we offer an integrated solution that uses the caption words to drive the feature grouping process. The result is a more general model acquisition framework that does not assume words correspond to individual features and does not require training images with isolated objects or unambiguous labels. The model that is learned performs well at labeling cluttered scenes in a set of test images.

1. Introduction

Image annotation is recognized as an important means for associating meaning (in the form of caption words or keywords) with an image; it can also be seen as a means for assigning meaning (in the form of visual features) to the caption words (*e.g.*, [2, 9, 14, 16]). The patterns of co-occurrence of words and visual features in annotated images can provide the evidence needed to establish meaningful links between the visual and linguistic representations. However, this approach can only succeed to the extent that the words and visual features correspond to meaningful aspects of what is portrayed in the image.

On the language side, we face three problems. Words in the caption may be noisy (*e.g.*, misspelled), they may be irrelevant (*i.e.*, they don't refer to objects in the image), or they may, in isolation, not capture the best meaning of the object (*e.g.*, if, for a particular object in an image, "rocket ship" is more appropriate than "rocket" or "ship"). In this work, we will focus on only one of these problems, namely the problem of irrelevant words in the caption. On the vision side, we will use the associations between caption words and image features to overcome all three analogous problems in vision, eliminating unstable (noisy) features from a model, excluding background (irrelevant) features from the model, and grouping individual features belonging to an object into collections that better capture the scope (granularity) of the object.

A cluttered scene may yield hundreds or even thousands of local features, only a small subset of which corresponds to any given object. This perceptual grouping or segmentation problem exists no matter what type of image representation is used: pixels, line segments, local features or regions. Given a set of features extracted from an image, the challenge is to find the meaningful subsets, to 'carve nature at its joints.' However, these 'important' subsets represent a vanishingly small portion of all possible subsets for any non-trivial representation. Although effective bottom-up grouping heuristics exist for certain classes of features (*e.g.*, Gestalt grouping of lines), today's popular interest point-based features do not lend themselves to bottom-up grouping. Simply evaluating all possible groupings is not usually feasible.

How might we find meaningful groupings in the context of the annotation problem? Simple frequency of occurrence in the training data can provide a clue. If certain collections of image features exist more often than can easily be explained by chance, they may have a common, meaningful source. However, the number of such groupings may still be prohibitively large. Even if they do arise from some common, recurring source in the image, such groupings might have no corresponding word on the language side. This suggests a dual approach: find co-occurring visual features which *also* have a significant level of co-occurrence with specific words. Considering the modalities of language and vision together can make the perceptual grouping problem more tractable at the same time as it offers a solution to the semantic association problem.

In this paper, we begin with a set of captioned training images of cluttered scenes containing multiple objects. On the

vision side, each image is processed to yield a set of local SIFT features [13], yielding hundreds or thousands of features per image. On the language side, each image is annotated with a set of nouns which may or may not name objects in the image. Drawing on the probabilistic translation model of [4] (as in [9, 16]), we introduce a novel iterative algorithm for growing candidate associations between individual SIFT features and nouns into more definitive models of object appearance in the form of collections of SIFT features, or *compounds*. This simultaneous language-driven perceptual grouping and association yields a set of models which is subsequently used to annotate new images.

2. Related Work

A number of researchers have explored the problem of learning associations between image features and text, including Barnard and Forsyth [2], Duygulu *et al*. [9], Blei and Jordan [3], and Cascia *et al*. [8]. As impressive as the results are, these approaches make limiting assumptions that prevent them from associating words with *configurations* of features (though see Fergus *et al*. [11] and Wachsmuth *et al*. [16]).

Cascia *et al*. compute visual features over the whole image and therefore do not support the notion of an image object distinct from its background. Other approaches, like Barnard and Forsyth [2], Feng *et al*. [10], or Carneiro and Vasconcelos [7], divide the image into a pre-segmented set of regions or rectangular image tiles and assume probabilistic generation processes in which words and image features are independently produced given the image or image topic. Thus, words are never directly linked to groups of image features.

Though the approaches by Feng *et al*. or Carneiro and Vasconcelos implicitly capture some compound knowledge by coding positional information in the different image tiles, the representation is not explicit. Models proposed by Duygulu *et al*. [9] as well as Blei and Jordan [3] include an explicit alignment of caption words and image regions which is a prerequisite for compound modeling. However, words are only aligned with individual regions and not configurations. In Carbonetto *et al*. [5], relations between regions are included in a Markov random field model that could, in principle, capture the interrelations between parts.

Acknowledging that coarse granularity features, such as regions, may be oversegmented during feature extraction, Barnard *et al*. [1] proposed a ranking scheme for potential merges of regions based on a model of similar word-region association. This approach is problematic, however, as disparate components of a compound may have very different word associations.

Wachsmuth *et al*. [16] proposed a framework employing a translation model to help extract shape categories of common object classes from collections of annotated, oversegmented images. While performing perceptual grouping through region merges, the framework would also attempt to detect object classes composed of multiple parts. The technique, however, was only demonstrated on synthetic images, and training focused on finding the correct (possibly hierarchical) image segmentations, rather than distinctive configurations of features.

In a different context, Hoogs *et al*. [12] used co-occurring visual features to help suggest a semantic interpretation of the image content. Detected image regions activate elemental terms in WordNet which, together with topics extracted from accompanying text, help guide the semantic search through the knowledge base.

In contrast to the automatic image annotation literature, object recognition techniques often employ models describing multiple components and their relationships. As a relatively recent example, Fergus *et al*. [11] model objects as constellations of salient features. Their approach, however, does not explicitly deal with the ambiguity of training sets containing multiple objects and multiple noisy annotation words for each image.

3. Features and Compound Features

In an image annotation system, the choice of feature largely determines the types of object classes that can be reliably detected. Some object types, such as grass, pavement or sky, have no consistent shape, and might be well-described by a region descriptor (*e.g*., a blob) encoding color or texture. Other objects, such as tables, lamps and clothed people, are defined more by their shape and less by their color and texture, suggesting a structured or parameterized shape model. Still other object classes, such as trees or mountains, exhibit patterns of limited variation in both shape and appearance.

In this work, we adopt the local interest point detector and SIFT feature representation developed by Lowe [13]. Briefly reviewing, a set of interest points are detected at the scale-space maxima of an image; let p_j represent the jth detected local interest point with a particular position, orientation and scale. A SIFT feature, $\mathbf{x}_j$, encodes the local pattern of intensity changes as a 128-element vector. The feature is invariant to scaling, translation and rotations in the image plane, and is designed to be robust to changes in intensity and contrast, small translation errors, and small rotations in depth.

In adopting SIFT features, we restrict ourselves to object classes which generate interest points in reasonably stable configurations and where the surface appearance is stable for objects within the class. For most purposes, this means exemplar objects or objects manufactured to the same design. In cluttered scenes, SIFT features are plentiful, with thousands detectable in a typical image. This is both a blessing and a curse. It allows us to potentially detect objects that occupy only a small portion of the image, yet the number of visual features far outweighs the number of relevant words

in a typical image annotation.

The SIFT feature vectors $\mathbf{x}$ are continuous, while our current translation model operates on discrete tokens. We therefore perform vector quantization to replace each feature $\mathbf{x}_j$ with a discrete 'visual word' feature, v_j. Other recently-developed image annotation techniques avoid this quantization step and estimate continuous probability densities over a feature space [7, 10]. Since we are employing SIFT features, which already have high dimensionality, and wish to find collections or configurations of these features, the continuous description would probably become too unwieldy. Of course, transforming the feature vectors into discrete classes introduces an unavoidable level of quantization noise. It may be possible to reduce the effect of this noise by associating a small weighted set of features with each interest point.

Vector quantization is trained on a set of $300,000$ local interest features selected at random from a pool of 2500 stock photo images. We use the K-means algorithm to generate a set of $V = 5000$ cluster centers, $\bar{\mathbf{x}}_v$, $v \in \{1, \ldots, V\}$, similar to the approach of Sivic and Zisserman [15]. The feature data is whitened before clustering so that the Euclidean distance on the transformed data equals the Mahalanobis distance in the original space:

$$d(\mathbf{x}_j, \mathbf{x}_v) = \sqrt{(\mathbf{x}_j - \bar{\mathbf{x}}_v)^T \Sigma^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}_v)} \qquad (1)$$

where Σ is the covariance matrix of the interest features. Each detected SIFT feature is replaced with the index of the nearest cluster center:

$$v_j = \arg\min_v \; d(\mathbf{x}_j, \bar{\mathbf{x}}_v). \qquad (2)$$

At training time, we also calculate the proportion α_v of SIFT features in our stock photo collection assigned to each cluster center $\bar{\mathbf{x}}_v$. There is a great deal of variation in these cluster weights. Again following Sivic and Zisserman [15], we suppress the most common 0.5 percent and the least-common 10 percent of features. This is based on an analogy with text retrieval, where the least common and most common terms are less informative.

Individual SIFT features may not be strong indicators of a particular object or object class. Specific arrangements or structures of local features are more discriminative. In this paper, we consider features that exist within a local neighborhood. A *compound feature* c_m is essentially a 'bag' of such local features. Each compound feature is a triple, $c_m = \{\mathcal{V}_m, \eta_m, k_m\}$, consisting of a set of visual features $\mathcal{V}_m$, a presence threshold η_m, and a neighborhood size k_m. A compound is considered present if at least η_m distinct features within $\mathcal{V}_m$ exist within a 'local neighborhood' of size k_m. Specifically, consider an interest point p_j with the corresponding SIFT feature $v_j \in \mathcal{V}_m$. The feature v_j and the features of the k_m spatially closest interest points form the neighborhood feature set $\mathcal{N}_j^{k_m}$. We detect the compound c_m at p_j if at least η_m distinct elements of $\mathcal{V}_m$ are present in $\mathcal{N}_j^{k_m}$. We ignore compound features whose neighborhood would overlap with a previously detected instance of the same compound. Note that each individual feature can also be considered a compound of size one, so $\mathbf{c}$ is the entire set of local features and detected compounds within an image. The combinatorics of the problem yields a very large number of possible compound features in an image. However, only a vanishingly small fraction of these correspond to meaningful configurations, and fewer still may correspond to nouns appearing in the captions. This further increases the asymmetry between annotation word counts and visual feature counts.

4. The Caption-to-Image Translation Model

In this work, the translation model serves two purposes. Given a set of images, each of which is associated with a set of visual features and a collection of annotation words, the translation model discovers correspondences between the visual features and the annotations. However, the initial set of visual features, extracted based on image characteristics alone, may not be distinctive enough to pick out the objects that are named in the annotations. Therefore we also use the translation model to help guide a search for better features. An iterative process proposes larger, more distinctive compound features, and uses the correspondence strength in the current translation model to evaluate the goodness of each potential compound. In the following subsections, we discuss each of these aspects of the translation model in turn.

4.1. The Basic Translation Model

As in earlier image annotation work (e.g., [9, 16]), we begin with the following translation model: the conditional probability $Pr(\mathbf{f}|\mathbf{e})$, given two sets of symbols $\mathbf{f}$ and $\mathbf{e}$. (In the formulation by Brown *et al*. [4] for machine translation, $\mathbf{f}$ and $\mathbf{e}$ referred to sequences of words in French and English, respectively.) To reduce the number of parameters to be estimated, it is generally assumed that the symbols (words or image features) can be generated independently. Each symbol f_i has an alignment variable a_i from $\{0, \ldots, L\}$, where L is the number of symbols in $\mathbf{e}$, that associates f_i with a single symbol, e_{a_i}, which may be the "null" symbol, e_0.

There is a strong asymmetry in this model, as each symbol in $\mathbf{f}$ is associated with a single symbol in $\mathbf{e}$ (or with none of those—the null symbol), while each symbol in $\mathbf{e}$ can be associated with an arbitrary number of symbols in $\mathbf{f}$. In our annotated images, we also have an asymmetry, in that there are typically a very large number of visual features and a relatively small number of caption words. It is much more likely that multiple visual features correspond to a single word than vice versa. We thus treat the set of annotation words $\mathbf{w}$ as $\mathbf{e}$ and the set of visual features $\mathbf{c}$ as $\mathbf{f}$ in the formula above, yielding the following, based on Brown *et al*.'s

Model 1:

$$Pr(\mathbf{c}|\mathbf{w}) = \frac{\epsilon}{(L+1)^M} \prod_{j=1}^{M} \sum_{a_j=0}^{L} t(c_j|w_{a_j}) \qquad (3)$$

Here M is the number of compound features, L is the number of caption words, ϵ is a constant, and $t(c_j|w_{a_j})$ is an element of the translation table $\mathbf{t}$ defining the distribution over compound visual features for each word. As in Brown *et al*., we use EM to find the $\mathbf{t}$ that maximizes the probability of the training data.

4.2. Dealing with "Noisy" Features

The goal is to learn stable associations between words and image features. One of the drawbacks of the above model is that it is devised for a situation—translation between two languages—in which most elements in one representation (the source language) are aligned with an element in the other representation (the target language). The possibility of alignment with the null word exists, but most words are expected to align with an actual word. However, this is generally not the case in aligning visual features and caption words, and is especially not the case with SIFT features. There are thousands of SIFT features in any given image, and only those that are a stable part of the appearance of a named object have any counterpart in the caption text. This motivates a larger role for the null symbol, to serve as a "default" alignment for the many SIFT features which correspond to objects or surfaces that are not named among the annotation words, or that are transient, unstable aspects of a named object. We want to ensure that such "chance" features are not linked to actual caption words, by increasing the likelihood that they align to the null word.

One issue is that our small pool of labeled training images is not broad enough for us to determine the distribution of the types of background features in images more generally. To address this, we use the counts of singleton features over our pool of 2500 stock images (see Section 3) to estimate the prior probability distribution over individual features. The prior likelihood of any compound feature is calculated assuming singleton features are placed independently. We then add a dummy entry to the translation training data which includes a set of visual feature counts (both singleton and compound) in proportion to their prior likelihoods (effectively estimating their occurrence by chance). The dummy entry has no associated caption words, entailing that a strong association is established between the "chance" features and the null word. Then, features with high prior likelihood or which appear rarely in the training images are more likely to align with the null symbol (*i.e*., to be considered background).

Since the background or noisy features typically comprise more of the image than features from the objects of interest, the translation table is normalized to give the null word a higher alignment probability than the actual caption words. In our experiments, visual features are ten times more likely *a priori* to align with the null word than with a caption word.

4.3. The Search for Compound Features

A cornerstone of our framework is the use of the associations between annotation words and visual features to guide the process of grouping visual features into meaningful collections that serve as better indicators of the objects named by the words. This is achieved by initially learning the translation probabilities on singleton visual features, then iteratively trying out potential collections of the existing features. These potential compound features are evaluated with respect to their improvement to the translation probability for predicting the annotation word under consideration. A potential feature that leads to an improvement is adopted, and the translation model is iteratively re-learned.

We use a simple greedy algorithm. We start by initializing the set of features that will be considered. For each of the W annotation words w_i (excluding the null word), we choose the $n_{seed} = 20$ singleton visual features v which have the highest $Pr(w_i|v)$ and occur more than $n_{min} = 8$ times in the training data. These features form $C = \{c_{ij}|i = 1 \ldots W, j = 1 \ldots n_{seed}\}$, the basis for the parallel, independent growth of compound features. Initially, all c_{ij} are singleton features, while on successive iterations, they may be singleton or compound features.

Each iteration of the search algorithm considers all elements of C. If c_{ij} is the current feature, we try one previously untested modification to produce a new compound feature m_{ij} for the associated word w_i. The compound m_{ij} replaces its constituent features wherever they appear together in the training images. If m_{ij} occurs at least n_{min} times, we update the translation model for the altered data to get $Pr(w_i|m_{ij})$. If this is greater than $Pr(w_i|c_{ij})$, then m_{ij} replaces c_{ij} in C. Either way, we "undo" the changes to the training data and continue the iterative process with the next feature in C.

The possible modifications for producing potential compounds include removing any one element from the current compound, and adding any element that co-occurs (in the same neighborhood) with the compound in the training data. For each removal/addition operation, we test both a 'changed' and a 'stable' version where the target number η_{ij} of constituent features is decreased/increased by one, respectively, in the former, and unchanged in the latter.

In theory, we could run the search until there are no more potential changes in C. In practice, this is too time-consuming, and we limit the outer loop of the search to 200 iterations.

5. Experiments

5.1. Data Set

We tested the algorithm on a set of real object images, in this case 228 images of arrangements of children's toys. The original color photographs were converted to intensity images with a resolution of 800x600. Most images contain 3 or 4 toy objects out of a pool of 10, though there are a handful of examples of up to 8 objects. The objects are not arranged in any consistent pose and many are partially occluded. Illumination was either direct sunlight, indirect natural light or from the camera's integrated flash. The images were collected against approximately 15 different backgrounds of varying complexity.

The pool of 228 images was randomly divided into a training set of 128 and a test set of 100. Each training image was annotated with the unique keyword for each object of interest shown and between 2 and 5 other keywords uniformly drawn from a pool of distractor labels. Figure 1 displays some example images from the training set and their associated annotations. Note that the objects of interest never appear individually and the training data contains no information as to the position or pose of the labeled objects.

5.2. Results

We ran the compound feature search technique on the training set with all compounds set to a fixed neighborhood size of $k_m = 100$. Figure 2 illustrates some example instances in the training set of the single best compound feature for the each of the 'rocket', 'ernie', 'horse' and 'bug' objects. Locations of SIFT features that form the detected compound are marked with yellow circles while a blue rectangular region indicates the approximate extent of the local neighborhood.

Note that the spatial configuration of component features sometimes varies across detections of the same compound. In some cases, a compound may even match more than one location on the same object. This degree of flexibility allows the compound to compensate for noise in the feature extraction process and to match the object across orientation changes and occlusions. However, such a pliable configuration definition is also more likely to generate false detections.

Once we have generated a set of learned compound features C from the training set, we employ a simple technique to annotate test images. Given a confidence threshold, Δ, we remove from C all learned compound features c_{ij} where $Pr(w_i|c_{ij}) < \Delta$. On loading a new test image, the algorithm extracts a set of SIFT features and their associated local neighborhood structure. If the image contains one or more instances of a compound c_{ij} in C, we label the image with the corresponding word, w_i. Labels are binary, as multiple compound detections do not necessarily indicate multiple instances of the object.

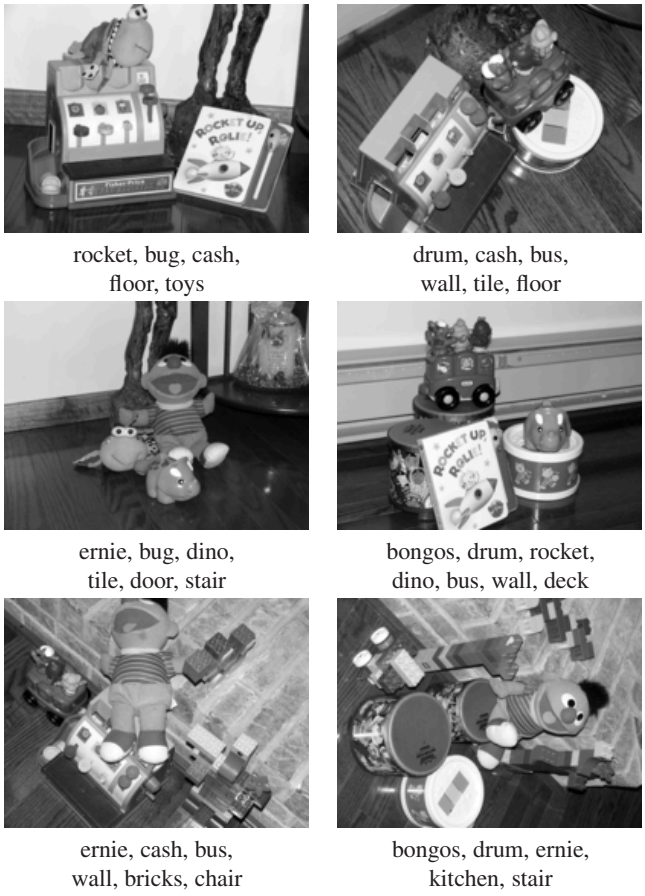
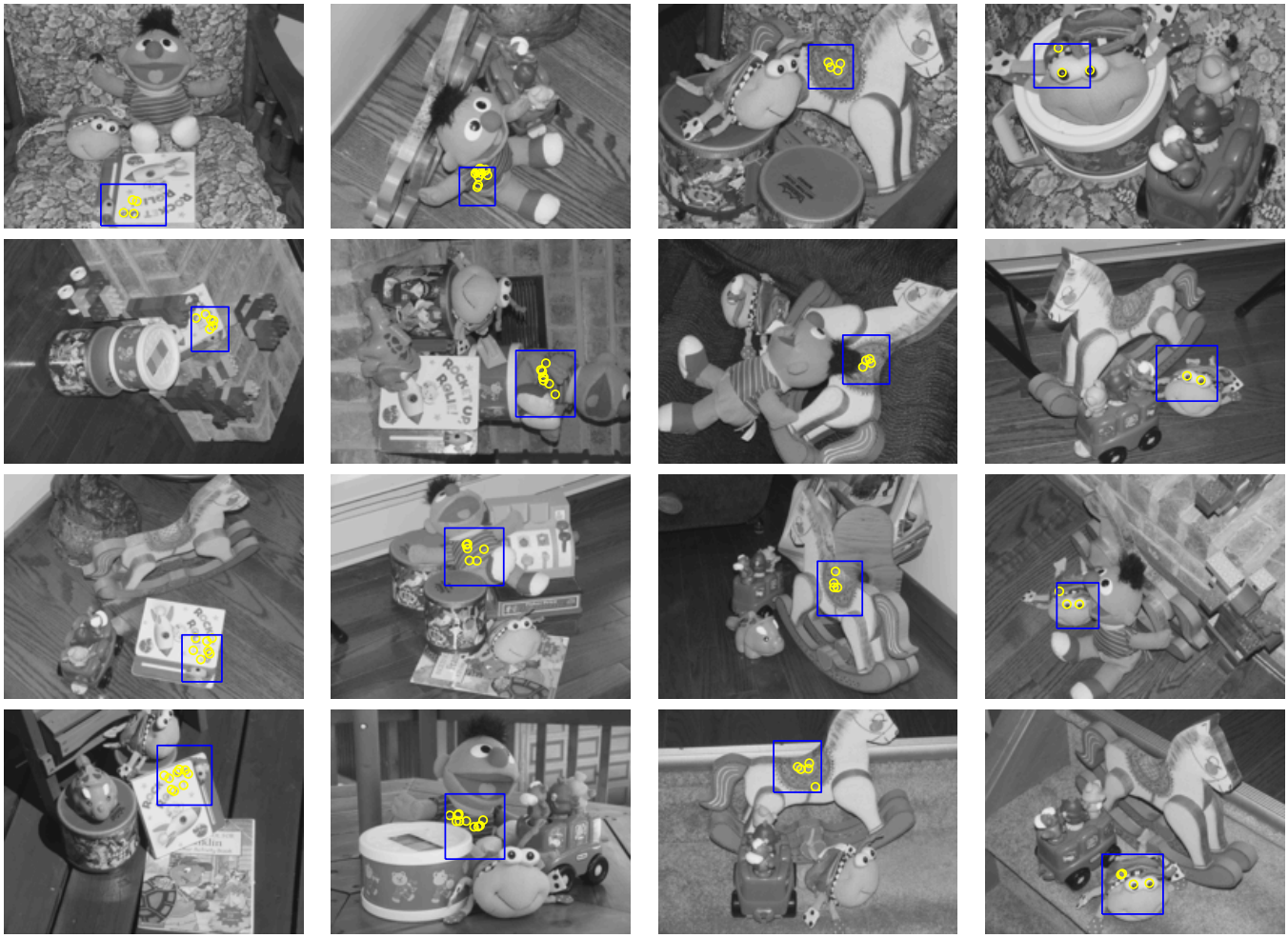


Figure 1. Example training images and their associated labels. Approximately half the labels (*e.g.* 'wall', 'tile', 'floor') serve as distractors for the relevant labels such as 'rocket', 'bug' and 'cash'.

Precision is the portion of detected annotations that are correct, while recall is the proportion of correct annotations that are properly detected. The precision-recall curves in Figures 3 and 4 represent the output with the confidence threshold Δ ranging from 1 to 0. Figure 3 demonstrates that the feature compounds learned on the training set are more distinctive and a stronger basis for annotation than the individual features. On the whole, single SIFT features do not perform substantially better than chance using this simple annotation scheme.

The fact that the curves are not monotonic indicates that confidence on the training set is not always a good predictor of performance on the test set. In fact, a few compounds with a moderately high confidence score in training appear to encode patterns that do not fall on the object of interest. However, most compounds that result from such coincidental correlations in the training data have low confidence scores.

Figure 4 shows the precision recall curves broken down by individual object. Results are fairly good for several objects, notably the 'rocket' and 'cash' exemplars. Given the variety of object poses and prevalence of partial occlusions,



(a) (b) (c) (d)

Figure 2. Example detections of compound features associated with the labels (a) ‘rocket’, (b) ‘ernie’, (c) ‘horse’ and (d) ‘bug’, respectively.

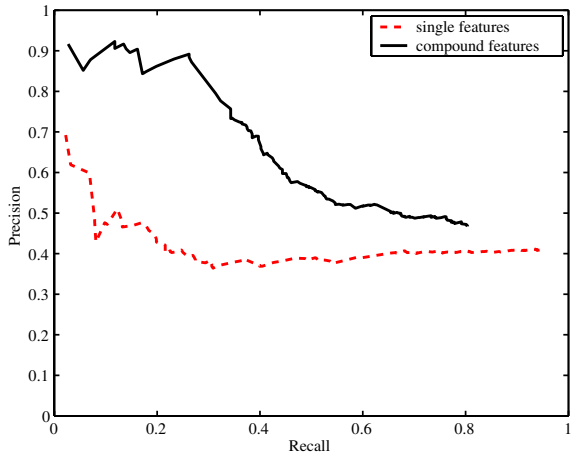
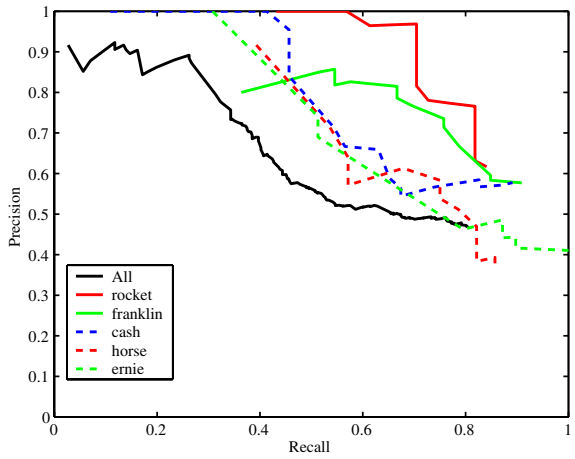


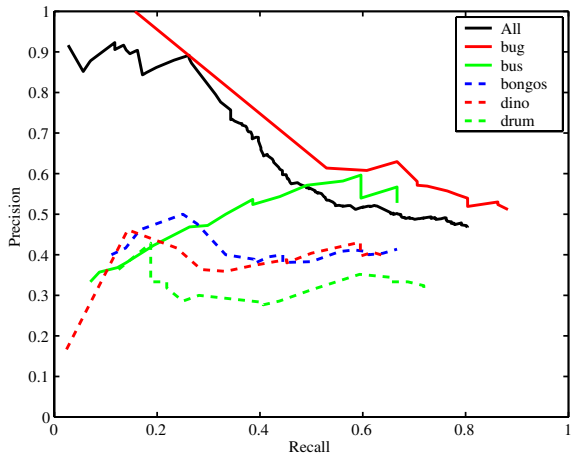
Figure 3. Precision vs. recall for single features versus compound features on the test set. Single features are not as distinctive.

even the stronger compounds learned in training do not detect every instance of the object.

For some objects, most notably ‘dino’, the system fails to find any useful identifying compounds. This is because none of the 20 original most highly-correlated seed features for the word ‘dino’ actually fall on the object of interest. Each is a relatively rare noise feature that happens to correlate with ‘dino’ more than any other label. This illustrates an important drawback of the current implementation: the object must have a feature that is somewhat distinctive in itself. It is likely that ‘dino’ displays reasonably stable common configurations of interest points, but if there is no single interest point that is relatively exclusive to the object of interest then noise features may dominate the seed set. To combat this effect, the algorithm could consider more potential starting points. In some cases, it might be necessary to start the language iteration above the single-feature level.



(a)



(b)

Figure 4. Precision-recall curves arranged into objects with better responses (a) versus weaker responses (b). The search process found fewer high-confidence compounds for the objects listed in (b).

Table 1 contains precision and recall results for two values of the confidence threshold, Δ. A high confidence threshold generally leads to relatively high precision in the annotation, though two compounds associated with the 'bongos' and 'franklin' labels have anomalously high probabilities, leading to relatively low precision on the test set for these two labels.

Finally, Figure 5 displays the annotation results using $\Delta = 0.85$ for a few randomly-selected test images. The results indicate that the proposed method is capable of correctly assigning highly distinctive features to names of real objects even though the training set contains no instances of the object or object name in isolation.

	$\Delta = 0.9$		$\Delta = 0.8$	
Label	Prec.	Rec.	Prec.	Rec.
bongos	0.42	0.14	0.50	0.25
franklin	0.77	0.70	0.71	0.76
drum	1.00	0.00	1.00	0.00
ernie	1.00	0.31	0.71	0.51
rocket	0.97	0.70	0.78	0.73
bug	1.00	0.16	0.61	0.61
cash	1.00	0.41	0.95	0.46
dino	1.00	0.00	1.00	0.00
bus	1.00	0.00	1.00	0.00
horse	0.92	0.39	0.71	0.54
Overall	0.87	0.27	0.71	0.38

Table 1. Precision-Recall results for two confidence thresholds.

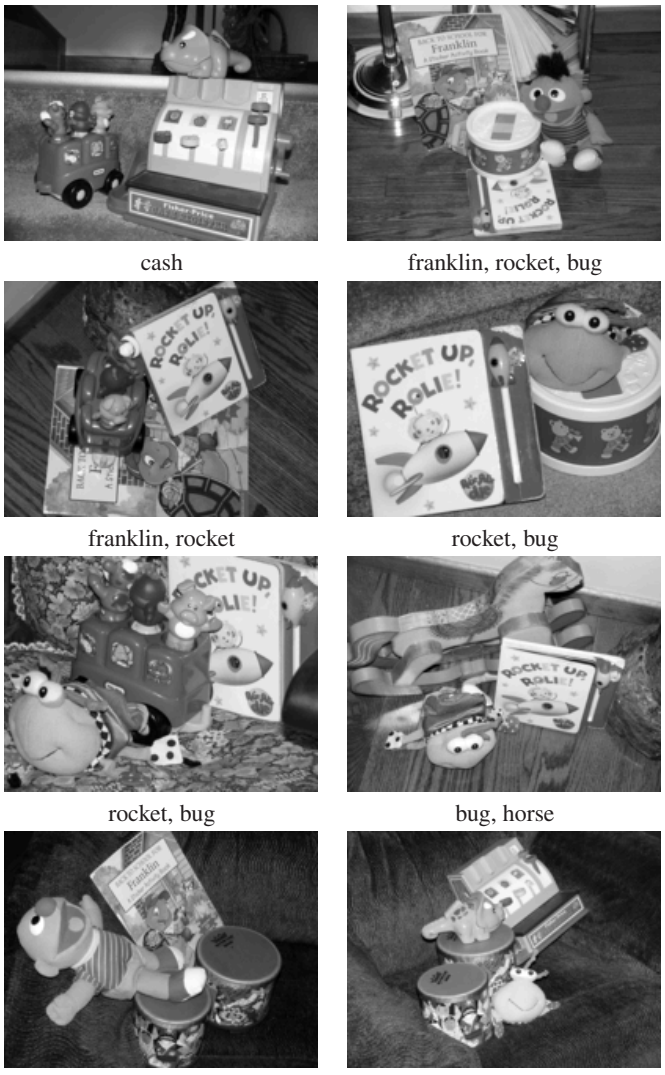


Figure 5. Randomly-selected test images with detected labels for $\Delta = 0.85$.

6. Discussion and Future Work

The method we have described can find groupings of SIFT features that are distinctive to individual objects and at the same time associate these objects with appropriate words from the image annotations. The system can detect these compounds and word correlations even though the features of interest themselves provide no grouping hint and always appear intermixed with features from other objects and complex backgrounds.

The system was implemented as a relatively simple vehicle to explore language-assisted grouping. As a practical image annotation system there are many directions for improvement. For instance, we would like to employ insights from computational linguistics to both construct a richer translation model and to allow us to move beyond individual words to groupings of words (compound nouns, collocations, or modifier-noun relations) that correspond to distinct visual patterns.

On the vision side, our current compound features, though simple and flexible, work over a limited range of scales and have very weak spatial constraints. An approach based on pairwise connections, such as that proposed by Carneiro and Jepson [6], could better model entire flexible or articulated objects while achieving much tighter spatial configuration constraints.

In both the visual and language domains, these more constrained compounds are more difficult to discover by brute force. We must induce them by exploiting patterns of co-occurrence both within and between the language and vision domains. For instance, a strong correlation in the training data between two or three SIFT features might provide a promising starting point to grow models of objects with no individually distinctive features.

Though SIFT features provide a strong basis for detecting unique objects, they are less ideal for detecting object categories or general settings. However, the grouping problem persists for most types of visual features and most forms of annotation, and many of the mechanisms for addressing the grouping problem for SIFT features are in fact quite general.

Any system for finding meaningful correlations between images and words also faces the problem of finding the level of description at which meaningful correlations exist. We expect that, in general, many correlations will exist not between individual words and individual features or regions, but between groups of words and collections of features. Patterns of correlation between the vision and language aspects of a data set provide important cues to meaningful grouping in both domains.

7. Acknowledgements

The authors gratefully acknowledge the support of Idée Inc., CITO, NSERC, and PREA.

References

[1] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth. The effects of segmentation and feature choice in a translation model of object recognition. In *CVPR*, pages 675–682, 2003.

[2] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *ICCV*, pages 408–415, 2001.

[3] D. M. Blei and M. I. Jordan. Modeling Annotated Data. Technical report, Computer Science Division, University of California, Berkeley, USA, 2002.

[4] P. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 32(2):263–311, 1993.

[5] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, pages 350–362, 2004.

[6] G. Carniero and A. Jepson. Flexible spatial models for grouping local image features. In *CVPR*, 2004.

[7] G. Carniero and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *CVPR*, 2005.

[8] M. L. Cascia, S. Sethi, and S. Sclaroff. Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web. In *Proc. of IEEE Workshop on Content-based Access of Image and Video Libraries*, June 1998.

[9] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, volume 4, pages 97–112, 2002.

[10] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[11] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

[12] A. Hoogs, J. Rittscher, G. Stien, and J. Schmiederer. Video content annotation using video analysis and a large semantic knowledgebase. In *CVPR*, 2003.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[14] D. Roy. Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication*, 4(1), 2002.

[15] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[16] S. Wachsmuth, S. Stevenson, and S. Dickinson. Towards a frameword for learning structured shape models from text-annotated images. In *HLT-NAACL03 Workshop on Learning Word Meaning from Non-Linguistic Data*, 2003.