Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data, pp. 22-29.

Towards a Framework for Learning Structured Shape Models from Text-Annotated Images

Sven Wachsmuth[+*], **Suzanne Stevenson**[+], **Sven Dickinson**[+]
[*] Bielefeld University, Faculty of Technology, 33594 Bielefeld, Germany
[+] University of Toronto, Dept. of Computer Science, Toronto, ON, Canada
{*swachsmu,suzanne,sven*}@*cs.toronto.edu*

Abstract

We present on-going work on the topic of learning translation models between image data and text (English) captions. Most approaches to this problem assume a one-to-one or a flat, one-to-many mapping between a segmented image region and a word. However, this assumption is very restrictive from the computer vision standpoint, and fails to account for two important properties of image segmentation: 1) objects often consist of multiple parts, each captured by an individual region; and 2) individual regions are often over-segmented into multiple subregions. Moreover, this assumption also fails to capture the structural relations among words, e.g., part/whole relations. We outline a general framework that accommodates a many-to-many mapping between image regions and words, allowing for structured descriptions on both sides. In this paper, we describe our extensions to the probabilistic translation model of Brown et al. (1993) (as in Duygulu et al. (2002)) that enable the creation of structured models of image objects. We demonstrate our work in progress, in which a set of annotated images is used to derive a set of labeled, structured descriptions in the presence of oversegmentation.

1 Introduction

Researchers in computer vision and computational linguistics have similar goals in their desire to automatically associate semantic information with the visual or linguistic representations they extract from an image or text. Given paired image and text data, one approach is to use the visual and linguistic representations as implicit semantics for each other—that is, using the words as names for the visual features, and using the image objects as referents for the words in the text (cf. Roy, 2002). The goal of our work is to automatically acquire structured object models from image data associated with text, at the same time learning an assignment of text labels for objects as well as for their subparts (and, in the long run, also for collections of objects).

Multimodal datasets that contain both images and text are ubiquitous, including annotated medical images and the Corel dataset, not to mention the World Wide Web, allowing the possibility of associating textual and visual information in this way. For example, if a web crawler encountered many images containing a particular shape, and also found that the word *chair* was contained in the captions of those images, it might associate the shape with the word *chair*, simultaneously indicating a name for the shape and a visual "definition" for the word. Such a framework could then learn the class names for a set of shape classes, effectively yielding a translation model between image shapes (or more generally, features) and words (Duygulu et al., 2002). This translation model could then be used to answer many types of queries, including labeling a new image in terms of its visible objects, or generating a visual prototype for a given class name. Furthermore, since figure captions (or, in general, image annotations) may contain words for entire objects, as well as words for their component parts, a natural semantic hierarchy may emerge from the words. For example, just as tables in the image may be composed of "leg" image parts, the word *leg* can be associated with the word *table* in a part-whole relation.

Others have explored the problem of learning associations between image regions (or features) and text, including Barnard and Forsyth (2001), Duygulu et al. (2002), Blei and Jordan (2002), and Cascia et al. (1998). As impressive as the results are, these approaches make limiting assumptions that prevent them from being appropriate to our goals of a structured

[0]Wachsmuth is supported by the German Research Foundation (DFG). Stevenson and Dickinson gratefully acknowledge the support of NSERC of Canada.

object model. On the vision side, each segmented region is mapped one-to-one or one-to-many to words. Conceptually, associating a word with only one region prevents an appropriate treatment of objects with parts, since such objects may consistently be region-segmented into a collection of regions corresponding to those components. Practically, even putting aside the goal of part-whole processing, any given region may be (incorrectly) oversegmented into a set of subregions (that are not component parts) in real images. Barnard et al. (2003) propose a ranking scheme for potential merges of regions based on a model of word-region association, but do not address the creation of a structured object model from sequences of merges. To address these issues, we propose a more elaborate translation/association model in which we use the text of the image captions to guide us in structuring the regions.

On the language side of this task, words have typically been treated individually with no semantic structure among them (though see Roy, 2002, which induces syntactic structure among the words). Multiple words may be assigned as the label to a region, but there's no knowledge of the relations among the words (and in fact they may be treated as interchangeable labels, Duygulu et al., 2002). The more restrictive goal of image labeling has put the focus on the image as the (structured) object. But we take an approach in principle of building a structured hierarchy for both the image objects and their text labels. In this way, we aim not only to use the words to help guide us in how to interpret image regions, but also to use the image structure to help us induce a part/whole hierarchy among the words. For example, assume we find consistently associated *leg* and *top* regions together referred to as a *table*. Then instead of treating *leg* and *table*, e.g., as two labels for the same object, we could capture the image part-whole structure as word relations in our lexicon.

Our goal of inducing associated structured hierarchies of visual and linguistic descriptions is a long-term one, and this paper reports on our work thus far. We start with the probabilistic translation model of Brown et al. (1993) (as in Duygulu et al., 2002), and extend it to structured shape descriptions of visual data. As alluded to earlier, we distinguish between two types of structured shape descriptions: collections of regions that should be merged due to oversegmentation versus collections of regions that represent components of an object. To handle both types, we incorporate into our algorithm several region merge operations that iteratively evaluate potential merges in terms of their improvement to the translation model. These operations can exploit probabilities over region adjacency, thus constraining the potential combinatorial explosion of possible region merges. We also permit a many-to-many mapping between regions and words, in support of our goal of inducing structured text as well, although here we report only on the structured image model, assuming similar mechanisms will be useful on the text side.

We are currently developing a system to demonstrate our proposal. The input to the system is a set of images segmented into regions organized into a region adjacency graph. Nodes in the graph encode the qualitative shape of a region using a *shock graph* (Siddiqi et al., 1999), while undirected edges represent region adjacency (used to constrain possible merges). On the text side, each image has an associated caption which is processed by a part-of-speech tagger (Brill, 1994) and chunker (Abney, 1991). The result is a set of noun phrases (nouns with associated modifiers) which may or may not pertain to image content. The output of the system is a set of many-to-many (possibly structured) associations between image regions and text words.

This paper represents work in progress, and not all the components have been fully integrated. Initially, we have focused on the issues of building the structured image models. We demonstrate the ideas on a set of annotated synthetic scenes with both multi-part objects and oversegmented objects/parts. The results show that at least on simple scenes, the model can cope with oversegmentation and converge to a set of meaningful many-to-many (regions to words) mappings.

2 Visual Shape Description

In order to learn structured visual representations, we must be able to make meaningful generalizations over image regions that are sufficiently similar to be treated as equivalent. The key lies in determining categorical shape classes whose definitions are invariant to within-class shape deformation, color, texture, and part articulation. In previous work, we have explored various generic shape representations, and their application to generic object recognition (Siddiqi et al., 1999; Shokoufandeh et al., 2002) and content-based image retrieval (Dickinson et al., 1998). Here we draw on our previous work, and adopt a view-based 3-D shape representation, called a shock graph, that is invariant to minor shape deformation, part articulation, translation, rotation, and scale, along with minor rotation in depth.

The vision component consists of a number of steps. First, the image is segmented into regions, using the mean-shift region segmentation algorithm of Comaniciu and Meer (1997).[1] The result is a region adjacency graph, in which nodes represent homogeneous

[1] The results presented in Section 4.2 are based on a synthetic region segmentation. When working with real images, we plan to use the mean-shift algorithm, although any region segmentation algorithm could conceivably be used.

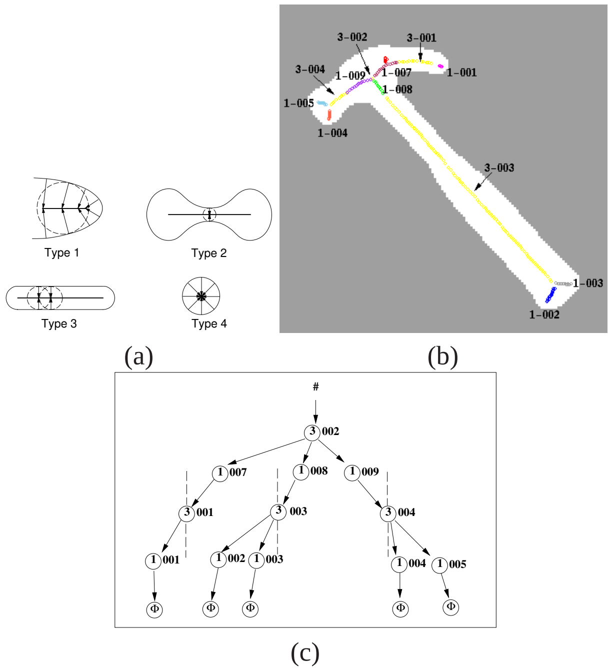


Figure 1: The Shock Graph Qualitative Shape Representation: (a) the taxonomy of qualitative shape parts; (b) the computed shock points of a 2-D closed contour; and (c) the resulting shock graph.

regions, and edges capture region adjacency. The parameters of the segmentation algorithm can be set so that it typically errs on the side of oversegmentation (regions may be broken into fragments), although undersegmentation is still possible (regions may be merged incorrectly with their neighbors). Next, the qualitative shape of each region is encoded by its shock graph (Siddiqi et al., 1999), in which nodes represent clusters of skeleton points that share the same qualitative radius function, and edges represent adjacent clusters (directed from larger to smaller average radii). As shown in Figure 1(a), the radius function may be: 1) monotonically increasing, reflecting a bump or protrusion; 2) a local minimum, monotonically increasing on either side of the minimum, reflecting a neck-like structure; 3) constant, reflecting an elongated structure; or 4) a local maximum, reflecting a disk-like or blob-like structure. An example of a 2-D shape, along with its corresponding shock graph, is shown in Figures 1(b) and (c).

The set of all regions from all training images are clustered according to a distance function that measures the similarity of two shock graphs in terms of their structure and their node attributes. As mentioned above, the key requirement of our shape representation and distance is that it be invariant to both within-class shape deformation as well as image transformation. We have developed

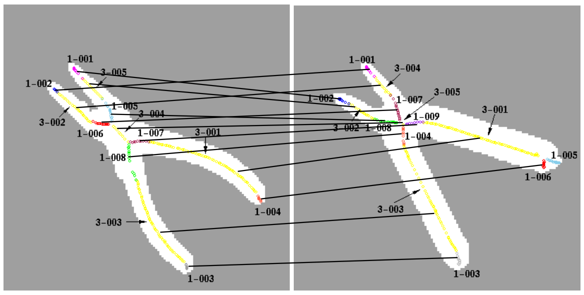


Figure 2: Generic Shape Matching

a matching algorithm for 2-D shape recognition. As illustrated in Figure 2, the matcher can compute shock graph correspondence between different exemplars belonging to the same class.

During training, regions are compared to region (shape) class prototypes. If the distance to a prototype is small, the region is added to the class, and the prototype recomputed as that region whose sum distance to all other class members is minimum. However, if the distance to the nearest prototype is large, a new class and prototype are created from the region. Using the region adjacency graph, we can also calculate the probability that two prototypes are adjacent in an image. This is typically a very large, yet sparse, matrix.

3 Learning of Translation Models

The learning of translation models from a corpus of bilingual text has been extensively studied in computational linguistics. Probabilistic translation models generally seek to find the translation string **e** that maximizes the probability $Pr(\mathbf{e}|\mathbf{f})$, given the source string **f** (where **f** referred to French and **e** to English in the original work, Brown et al., 1993). Using Bayes rule and maximizing the numerator, the following equation is obtained:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}} Pr(\mathbf{f}|\mathbf{e}) Pr(\mathbf{e}). \tag{1}$$

The application of Bayes rule incorporates $Pr(\mathbf{e})$ into the formula, which takes into account the probability that $\hat{\mathbf{e}}$ is a correct English string.

$Pr(\mathbf{f}|\mathbf{e})$ is known as the *translation model* (prediction of **f** from **e**), and $Pr(\mathbf{e})$ as the *language model* (probabilities over **e** independent of **f**). Like others (Duygulu et al., 2002), we will concentrate on the translation model; taking **f** as the words in the text and **e** as the regions in the images, we thus predict words from image regions. However, we see the omission of the language model component, $Pr(\mathbf{e})$ (in our case, probabilities over the "language" of images—i.e., over "good" region associations), as a shortcoming. Indeed, as we see below, we insert some simple aspects of a "language model" into our current formulation, i.e. using the region adjacency graph to restrict possible merges, and using the *a priori* probability of a region $Pr(r)$ if translating from words to regions. In future work, we plan to elaborate the $Pr(\mathbf{e})$ component more thoroughly.

Data sparseness prevents the direct estimation of $Pr(\mathbf{f}|\mathbf{e})$ (which predicts one complete sequence of symbols from another), so practical translation models must make independence assumptions to reduce the number of parameters needed to be estimated. The first model of Brown et al. (1993), which will be used and expanded in our initial formulation, uses the following approximation to $Pr(\mathbf{f}|\mathbf{e})$:

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} Pr(M) \prod_{j=1\ldots M} Pr(a_j|L) Pr(f_j|a_j, e_{a_j}) \quad (2)$$

where M is the number of French words in $\mathbf{f}$, L is the number of English words in $\mathbf{e}$, and $\mathbf{a}$ is an *alignment* that maps each French word to one of the English words, or to the "null" word e_0. $Pr(M) = \varepsilon$ is constant and $Pr(a_j|L) = 1/(L+1)$ depends only on the number of English words. The conditional probability of f_j depends only on its own alignment to an English word, and not on the translation of other words f_i. These assumptions lead to the following formulation, in which $t(f_j|e_{a_j})$ defines a translation table from English words to French words:

$$Pr(\mathbf{f}|\mathbf{e}) = \frac{\varepsilon}{(L+1)^M} \prod_{j=1\ldots M} \sum_{a_j=0\ldots L} t(f_j|e_{a_j}) \quad (3)$$

To learn such a translation between image objects and text passages, it is necessary to: 1) Define the vocabulary of image objects; 2) Extract this vocabulary from an image; 3) Extract text that describes an image object; 4) Deal with multiple word descriptions of objects; and 5) Deal with compound objects consisting of parts. Duygulu et al. (2002) assume that all words (more specifically, all nouns) are possible names of objects. Each segmented region in an image is characterized by a 33-dimensional feature vector. The vocabulary of image objects is defined by a vector quantization of this feature space. In the translation model of Brown et al., Duygulu et al. (2002) substitute the French string $\mathbf{f}$ by the sequence $\mathbf{w}$ of caption words, and the English string $\mathbf{e}$ by the sequence $\mathbf{r}$ of regions extracted from the image (which they refer to as blobs, $\mathbf{b}$). They do not consider multiple word sequences describing an image object, nor image objects that consist of multiple regions (oversegmentations or component parts).

In section 2 we argued that many object categories are better characterized by generic shape descriptions rather than finite sets of appearance-based features. However, in moving to a shape-based representation, we need to deal with image objects consisting of multiple regions (cf. Barnard et al., 2003). We distinguish three different types of multiple region sets:

1. Type A (accidental): Region over-segmentation due to illumination effects or exemplar-specific markings on the object that results in a collection of subregions that is not generic to the object's class.

2. Type P (parts): Region over-segmentation common to many exemplars of a given class that results in a collection of subregions that may represent meaningful parts of the object class. In this case, it is assumed that on some occasions, the object is seen as a silhouette, with no over-segmentation into parts.

3. Type C (compound): Objects that are always segmented into their parts (e.g., due to differently colored or textured parts). This type is similar to Type P, except that these objects never appear as a whole silhouette. (Our mechanism for dealing with these objects will also allow us, in the future, to handle conventional collections of objects, such as a set of chairs with a table.)

We can extend the one-to-one translation model in Eqn. (3) above by grouping or merging symbols (in this case, regions) and then treating the group as a new symbol to be aligned. Theoretically, then, multiple regions can be handled in the same translation framework, by adding to the sequence of regions in each image, the regions resulting from all possible merges of image regions:

$$Pr(\mathbf{w}|\mathbf{r}) = \frac{\varepsilon}{(\tilde{L}+1)^M} \prod_{j=1\ldots M} \sum_{a_j=0\ldots \tilde{L}} t(w_j|r_{a_j}) \quad (4)$$

where $\tilde{L}$ denotes the total number of segmented and merged regions in an image. However, in practice this causes complexity and stability problems; the number of possible merges may be intractable, while the number of semantically meaningful merges is quite small.

Motivated by the three types of multiple region sets described above, we have instead developed an iterative bootstrapping strategy that filters hypothetically meaningful merges and adds these to the data set. Our method proceeds as follows:

1. As in Dyugulu et al., we calculate a translation model $t^0(w|r)$ between words and regions, using a data set of N image/caption pairs $D = \{(\mathbf{w}_d, \mathbf{r}_d)|d = 1\ldots N\}$. $\mathbf{r}_d$ initially includes a region for each segmented region in image d.

2. We next account for accidental over-segmentations (Type A above) by adding all merges to the data set that increase the score based on the old translation model:

$$score(D^{i+1}) = \prod_{(\mathbf{w},\mathbf{r})\in D^{i+1}} P(\mathbf{w}|\mathbf{r}; t^i(w|r)) \quad (5)$$

That is, we use the current translation model to determine whether to merge any two adjacent regions into a new region. If the quality of the translation is improved by the merge, we add the new region to **r**. If the dataset was extended by any number of new regions, the algorithm starts again with step 1 and recalculates the translation model.

3. We then account for regular over-segmentation (Type P above) by extending the number of regions merged for adjacent region sets—i.e., merges are no longer restricted to be pairwise. In this step, though, only sets of regions that frequently appear together in images are candidates for merging. Again, those that increase the score are iteratively added to the data set until the data set is stable.

4. For compound objects (Type C above), the score criterion does not apply because the silhouette of the merged structure does not appear in the rest of the data set. Since the current translation model has no information about the whole object, merging the component regions cannot increase the quality of the translation model.

   Instead, we develop a new scoring criterion, based on Melamed (1997). First, the current translation model is used to induce translation links between words and regions, and the mutual information of words and regions is calculated, using the link counts for the joint distribution. Next, the increase in mutual information is estimated for a hypothetical data set D' in which the regions of potential compounds are merged. If a compound contributes to an increase in mutual information in D', then the merge is added to our data set.

5. The sequence of steps above is repeated until no new regions are added to the data set.

In our algorithm above, we mapped our three approaches to dealing with region merges to the three types of multiple regions sets identified earlier (Types A, P, C). Indeed, each step in the algorithm is inspired by the corresponding type of region set; however, each step may apply to other types. For example, in a given data set, the legs of a table may only infrequently be segmented into separate regions, so that a merge to form a table may occur in step 2 (Type A) instead of step 3 (Type P). Thus, the actual application of steps 2–4 depends on the precise make-up of regions and their frequencies in the data set.

In our demonstration system reported next, step 3 of the algorithm is currently applied without considering how frequent a region pair appears. It iteratively generates three pairwise merges, with the output restricted to those that yield a shape seen before. We expect that considering only frequent shape pairs will stabilize merging effects and reduce computational complexity for more expensive merge operations than on the synthetic dataset. Our implementation of step 4 is in the early stage and currently considers combinations of any two regions, whether adjacent or not. This causes problems for images with more than one object or additional background shapes.

4 Demonstration

4.1 Scene Generation

As this paper represents work in progress, we have only tested our model on synthetic scenes with captions. Scenes contain objects composed of parts drawn from a small vocabulary of eight shapes, numbered 1–8, in Figure 3. (Our shapes are specified in terms of qualitative relationships over lines and curves; precise angle measurement is not important, Dickinson et al., 1992.) To simulate undersegmentation, primitive parts may be grouped into larger regions; for example, an object composed of three parts may appear as a single silhouette, representing the union of the three constituent parts. To simulate oversegmentation, four of the shape primitives (1, 5, 6, 8) can appear according to a finite set of oversegmentation models, as shown in Figure 3. To add ambiguity, oversegmentation may yield a subshape matching one of the shape categories (e.g., primitive shape 5, the trapezoidal shape, can be decomposed into shapes 1 and 4) or, alternatively, matching a subshape arising from a different oversegmentation. For example, the shape in the bottom right of Figure 3 is decomposed into two parts, one of which (25, representing two parallel lines bridged at one end by a concave curve and at the other end by a line) occurs in a different oversegmentation model (in this case, the oversegmentation shown immediately above it).

Scenes are generated containing one or two objects, drawn from a database of six objects (two chairs, two tables, and two lamps, differing in their primitive decompositions), shown in Figure 4. Given an object model, a decomposition grammar (i.e., a set of rewrite rules) is automatically generated that takes the silhouette of the shape and decomposes it into pieces that are either: 1) unions of the object's primitive parts, representing an undersegmentation of the object; 2) the object's primitive parts; or 3) oversegmentations of the object's primitive parts. In addition, the scene can contain up to four background shapes, drawn from Figure 3. These shapes introduce ambiguity in the mapping from words to objects in the scene, and can participate in merges of regions in our algorithm. Finally, each scene has an associated text caption that contains one word for each database object, which specifies either the name of the whole object (*table*/*stand*, *chair*/*stool*, *lamp*/*light*), or a part of the ob-

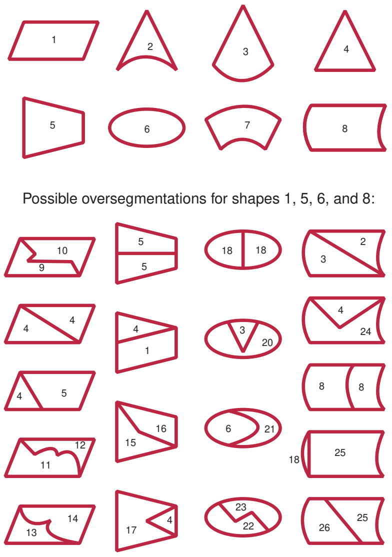


Figure 3: Top: The eight primitive shapes used to construct objects in the scene. Below: The various ways in which four of the shapes (1, 5, 6, 8) can be oversegmented.

ject (*base* or *leg*). Just as the scene contains background shapes, the caption may contain up to four “background” words that have nothing to do with the objects (or primitive parts) in the database.

We have developed a parameterized, synthetic scene generator that uses the derived rules to automatically generate scenes with varying degrees of undersegmentation, oversegmentation, ambiguous background objects, and extraneous caption words. Although no substitute for testing the model on real images, it has the advantage of allowing us to analyze the behavior of the framework as a function of these underlying parameters. Examples of input scenes it produces are shown in Figure 5.

4.2 Experimental Results

The first experiment we report here (Exp. 1) tests our ability to learn a translation model in the presence of Type A and Type P segmentation errors. We generated 1000 scenes with the following parameters: 1 or 2 objects per image, forced oversegmentation to a depth of 4, maximum 4 background shapes, one relevant word (part or whole descriptor), and maximum 2 meaningless random words per image. Table 1 shows the translation tables ($Pr(w|r)$) for this dataset, stopping the algorithm after step 1 (no merging) and after step 3. For all of the objects, the merging step increased the probability of one word, and decreased the probability of the others, creating a stronger word-shape association. For 5 of the objects, the highest probability word is a correct identifier of the object (*stand*, *chair*, *stool*, *light*, *lamp*), and for the

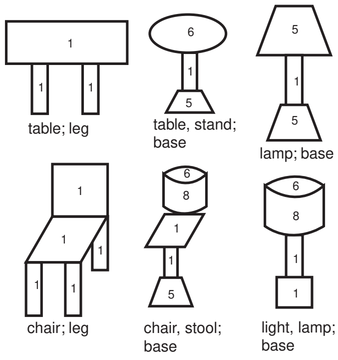


Figure 4: The database of six objects and associated words from which scenes are generated. Shape parts are labeled according to Figure 3.

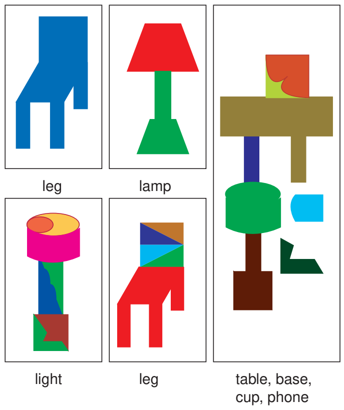


Figure 5: Examples of scenes input to our system.

other object, a word indicating a part of the object has high probability (*leg* for the first table object).

Although increasing the strength of one probability has an advantage, we need to explore ways to allow association of more than one “whole object” word (such as *lamp* and *light*) with a single object (cf. Duygulu et al., 2002). Since we maintain the component regions of a merged region, having both a part and a whole word, such as *leg* and *table*, associated with the same image is not a problem. Incorporating these into a structured word hierarchy should help to focus associations appropriately.

Another way to view the data is to see which shapes are most consistently associated with the meaningful words in the captions. Here we calculate $P(r|w)$ by $Pr(w|r)Pr(r)$, with the latter normalized over all shapes. A problem with this formulation is that, due to the $Pr(r)$ component, high frequency shapes can increase the probability of primitive components. However, the merging steps (2 and 3) of our algorithm raise the frequencies of complex (multi-region) shapes. Table 2 shows the five shapes with the highest values for each meaningful word, again before and after the merging steps in Exp. 1. Sev-

-	0.00	0.00	0.00	0.00	0.00	0.00
table	0.07	0.01	0.00	0.00	0.00	0.00
stand	0.00	0.98	0.00	0.00	0.00	0.00
chair	0.01	0.00	0.65	0.01	0.00	0.00
stool	0.00	0.00	0.00	0.13	0.00	0.00
lamp	0.00	0.00	0.00	0.00	0.41	0.80
light	0.00	0.00	0.00	0.00	0.33	0.00
leg	0.81	0.00	0.30	0.00	0.01	0.00
base	0.00	0.01	0.00	0.28	0.03	0.10

(a) before merging (step 1)

-	0.00	0.00	0.00	0.00	0.00	0.00
table	0.00	0.00	0.00	0.00	0.00	0.00
stand	0.00	1.00	0.00	0.00	0.00	0.00
chair	0.00	0.00	0.98	0.00	0.00	0.00
stool	0.00	0.00	0.00	0.41	0.00	0.00
lamp	0.00	0.00	0.00	0.00	0.03	0.99
light	0.00	0.00	0.00	0.00	0.97	0.00
leg	1.00	0.00	0.01	0.00	0.00	0.00
base	0.00	0.00	0.00	0.00	0.00	0.00

(b) after merging (steps 2/3)

Table 1: Exp. 1: Translation tables from shapes to words $P(w|r)$ for the 6 object silhouettes.

			6	1	
table	0.70	0.14	0.09	0.04	0.03
		6			1
stand	0.55	0.33	0.13	0.00	0.00
					1
chair	0.44	0.27	0.21	0.07	0.02
			8		
stool	0.32	0.31	0.24	0.06	0.06
	5				
lamp	0.64	0.10	0.08	0.07	0.07
					8
light	0.28	0.23	0.20	0.20	0.09
	1				
leg	0.84	0.15	0.01	0.00	0.00
base	0.28	0.23	0.11	0.11	0.09

Table 3: Exp. 2: The five shapes with highest $Pr(r|w)$ for the meaningful words, after step 4. Shape icons (for merged regions) or primitive shapes (indicated by number) have the probability for that word listed below.

eral complex shapes increase in probability after merging, and a number of new complex shapes appear in the lists.

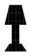
We report on one other experiment (Exp. 2) which was designed to test our approach to handling oversegmentations of Type C in step 4 of our algorithm. Our dataset again had 1000 images; here there was only one object per image, but every object was oversegmented into its primitive parts (that is, an object never appeared as a complete silhouette). (We did not allow oversegmentation of the primitives here, nor did we include irrelevant words in the captions.) Because our 6 objects never appear "whole," steps 2 and 3 of our algorithm cannot apply; before step 4, words are associated with primitive shapes only. After step 4, the highest probability word ($Pr(w|r)$) for 4 of the objects is a correct identifier of the object (*stand*, *chair*, *stool*, *light*); for one object, a word indicating a part of the object had high probability (*leg* for the rectangular table). (One object silhouette—the second lamp—was not fully reconstructed.) Table 3 shows the five shapes with the highest $Pr(r|w)$ values for each meaningful word, after step 4. For 3 of the whole object words (*stand*, *stool*, *light*), and both part words (*leg*, *base*), the best shape is a correct one. For the remaining whole object words (*table*, *chair*, *lamp*), a correct full silhouette is one of the top five. Step 4 clearly has high potential for reconstructing objects that are consistently oversegmented into their parts.

5 Conclusions

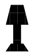
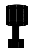
We have outlined a framework for the creation of associated visual and linguistic structured models, from images annotated with textual captions. Thus far, we have focused on the important open problem of dealing with oversegmentation in images. We have developed a set of extensions to a probabilistic translation model (Brown et al., 1993) that enable us to successfully merge oversegmented regions into coherent objects. Our initial experiments on synthetic data demonstrate that our algorithm can learn a useful translation model between image objects and words, even in the presence of substantial oversegmentation. We are currently experimenting with various parameters in our synthetic scene generator to guide further development of the algorithm, as well as experimenting on real data from the Web.

References

Steven Abney. 1991. Parsing by Chunks. In Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-Based Parsing*. Kluwer.

table	4 0.19	14 0.10	13 0.07	9 0.07	11 0.06
stand	18 0.24	 0.23	20 0.08	 0.08	16 0.06
chair	4 0.18	1 0.11	 0.09	24 0.08	13 0.06
stool	 0.16	 0.09	 0.09	 0.09	8 0.09
lamp	4 0.25	5 0.09	 0.07	10 0.06	16 0.06
light	18 0.16	25 0.08	22 0.06	2 0.06	21 0.06
leg	4 0.15	1 0.11	10 0.09	9 0.09	11 0.08
base	4 0.18	15 0.07	16 0.07	5 0.07	17 0.06

(a) before merging (step 1)

table	4 0.19	14 0.11	9 0.09	13 0.08	10 0.07
stand	 0.31	18 0.27	 0.11	20 0.08	19 0.05
chair	 0.19	4 0.14	 0.11	24 0.09	1 0.09
stool	 0.33	 0.16	 0.13	 0.11	3 0.05
lamp	4 0.25	5 0.20	 0.08	10 0.07	16 0.05
light	 0.29	 0.26	18 0.08	 0.08	6 0.04
leg	1 0.24	4 0.12	 0.08	10 0.08	9 0.08
base	4 0.20	15 0.08	16 0.08	22 0.06	17 0.06

(b) after merging (steps 2/3)

Table 2: Exp. 1: The five shapes with highest $Pr(r|w)$ for the meaningful words. Shape icons (for merged regions) or primitive shapes (indicated by number) have the probability for that word listed below.

Kobus Barnard and David Forsyth. 2001. Learning the Semantics of Words and Pictures. In *Proc. of Int. Conf. on Computer Vision (ICCV-2001)*, pages 408–415.

Kobus Barnard, Pinar Duygulu, Raghavendra Guru, Prasad Gabbur, and David Forsythi. 2003. The effects of segmentation and feature choice in a translation model of object recognition. In *Proc. of Computer Vision and Pattern Recognition*, page to appear.

David M. Blei and Michael I. Jordan. 2002. Modeling Annotated Data. Technical report, Computer Science Division, University of California, Berkeley, USA.

Eric Brill. 1994. Some advances in transformation-based part of speech tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence*, volume 1, pages 722–727, Menlo Park, CA, USA. AAAI Press.

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

Marco La Cascia, Saratendu Sethi, and Stan Sclaroff. 1998. 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.

D. Comaniciu and P. Meer. 1997. Robust analysis of feature spaces: Color image segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 750–755.

S. Dickinson, A. Pentland, and A. Rosenfeld. 1992. 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):174–198.

S. Dickinson, A. Pentland, and S. Stevenson. 1998. Viewpoint-invariant indexing for content-based image retrieval. In *IEEE International Workshop on Content-based Access of Image and Video Databases*, Bombay.

P. Duygulu, Kobus Barnard, J.F.G. de Freitas, and D.A. Forsyth. 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proc. of European Conference on Computer Vision (ECCV-2002)*, volume 4, pages 97–112.

I. Dan Melamed. 1997. Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'97)*, Providence, RI.

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

A. Shokoufandeh, S. Dickinson, C. Jonsson, L. Bretzner, and T. Lindeberg. 2002. The representation and matching of qualitative shape at multiple scales. In *Proceedings, ECCV*, pages 759–775, Copenhagen.

K. Siddiqi, A. Shokoufandeh, S. Dickinson, and S. Zucker. 1999. Shock graphs and shape matching. *International Journal of Computer Vision*, 30:1–24.