A Sentence is Worth a Thousand Pixels

Sanja Fidler
TTI Chicago
fidler@ttic.edu

Abhishek Sharma
University of Maryland
bhokaal@cs.umd.edu

Raquel Urtasun
TTI Chicago
rurtasun@ttic.edu

Abstract

We are interested in holistic scene understanding where images are accompanied with text in the form of complex sentential descriptions. We propose a holistic conditional random field model for semantic parsing which reasons jointly about which objects are present in the scene, their spatial extent as well as semantic segmentation, and employs text as well as image information as input. We automatically parse the sentences and extract objects and their relationships, and incorporate them into the model, both via potentials as well as by re-ranking candidate detections. We demonstrate the effectiveness of our approach in the challenging UIUC sentences dataset and show segmentation improvements of 12.5% *over the visual only model and detection improvements of* 5% *AP over deformable part-based models [8].*

1. Introduction

Images rarely appear in isolation. Photo albums are usually equipped with brief textual descriptions, while images on the web are usually surrounded by related text. In robotics, language is the most convenient way to teach an autonomous agent novel concepts or to communicate the mistakes it is making. For example, when providing a novel task to a robot, such as "pass me the stapler", we could provide additional information, e.g., "it is next to the beer bottle on the table". This textual information could be used to greatly simplify the parsing task. Despite this, the current most popular active learning paradigm is to provide the learner with additional labeled examples which were ambiguous or wrongly parsed, resulting in a tedious process.

In the past decade, we have witnessed an increasing interest in the computer vision community into leveraging text and image information in order to improve image retrieval [12, 27] or generate brief description of images [7, 15, 19, 1]. However, very few approaches [18, 26] try to use text to improve semantic understanding of images beyond simple image classification [21], or tag generation [6, 2]. This is perhaps surprising, as image descriptions

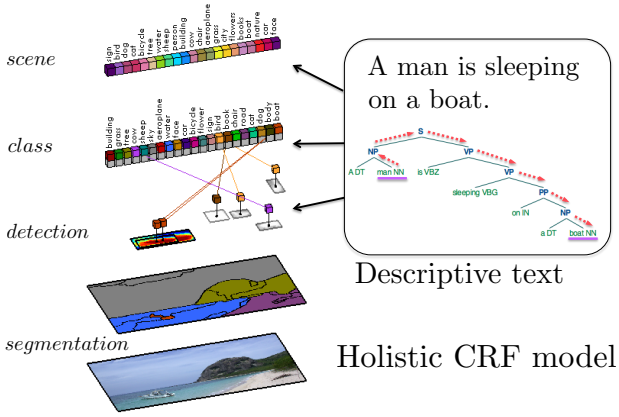


Figure 1. Our holistic model which employs visual information as well as text in the form of complex sentences.

can resolve a lot of ambiguities inherent to visual recognition tasks. If we were able to retrieve the objects and stuff present in the scene, their relations and the actions they perform from textual descriptions, we should be able to do a much better job at automatically parsing those images.

Here we are interested in exploiting textual information for semantic scene understanding. In particular, our goal is to reason jointly about the scene type, objects, their location and spatial extent in an image, while exploiting textual information in the form of complex sentential image descriptions generated by humans.

Being able to extract semantic information from text does not entirely solve the image parsing problem, as we cannot expect sentences to describe everything that is happening in the image, and in too great detail. Recent studies have shown that humans describe certain things and not others, perhaps as they are considered more important [3] or more surprising. Furthermore, not all information in the descriptions may be visually relevant, thus textual information also contains considerable noise for the task of interest, which needs to be properly handled. Moreover, natural language parsing algorithms are not perfect, resulting in noisy estimates.

In this paper we propose a holistic model for semantic parsing which employs text as well as image information.

Our model is a conditional random field (CRF) which employs complex sentential image descriptions to jointly reason about multiple scene recognition tasks. We automatically parse the sentences and extract objects and their relationships, and incorporate those into the model, both via potentials as well as by re-ranking the candidate bounding boxes. We demonstrate the effectiveness of our approach in the challenging UIUC dataset [7], which is a subset of PASCAL VOC 2008, and show that by employing textual information, our approach is able to improve detection's AP by 5% over deformable part-based models [8] and segmentation AP by 12.5% over state-of-the-art holistic models [29].

2. Related Work

In the past decade, a variety of approaches have looked into linking individual words to images, e.g., by predicting words from image regions [6, 2]. More recently, some approaches tried to generate sentential descriptions from images [7] and video [1, 9]. The generated sentences are semantically richer than a list of words, as the former describe objects, their actions and inter-object relations. [7] generated sentences by learning a meaning space which is shared between text and image domains. As the latent space is symmetric, images can also be retrieved given descriptions.

Recognition algorithms (e.g., object detectors) have been employed in order to produce more accurate sentence generation. The idea is to use the detected (tracked, in case of a video) objects and their estimated actions to produce richer sentences. In [15], detection and segmentation are exploited to construct a CRF that reasons about actions and object attributes. When a large dataset of images and associated captions is available, NN techniques can be employed to perform sentence retrieval. In [19], a few neighbors are retrieved using simple global descriptors. These neighbors are further re-ranked based on more semantic image content: objects, stuff, people and scene type. In [1], sentences were generated by parsing semantically videos into objects and their relations defined by actions.

Topic models have been widely employed to model text and images, particularly for retrieval tasks. Corr-LDA [4] was proposed to capture the relations between images and text annotations, but assumes a one to one correspondence between text and image topics. Different approaches have been proposed to avoid this assumption. In [20], regression is used to predict the topics in one modality from the topics in the other. This is further extended in [12] so that each image does not have to be associated with a text description. Topic models have also been used for word sense disambiguation [22], where the task is to learn a visual model based only on the name of an object. As words are generally polysemous, this approach can lead to noisy models if no special treatment of the different senses is performed.

Berg et al. [3] aimed at predicting what's important in images. They used sentences generated by humans describing the images in order to analyze what humans think is important. Their approach assumes that composition (e.g., size, location), semantics (e.g., object classes) as well as contextual factors (i.e., attribute-object and object-scene pairs) are a good proxy for predicting importance.

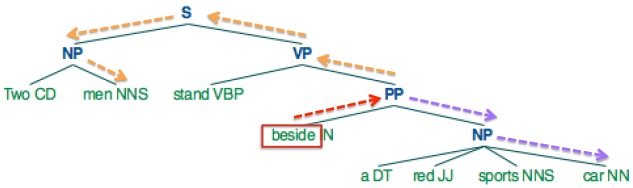


Figure 2. Extracting prepositional relations from text.

Despite the large body of work in generating descriptions or word annotations from images, there is little work into trying to use text to improve recognition. In [21], image captions were employed in a transfer learning scenario to improve image classification. A dataset of images and tags is employed in [27] to learn mutimodal deep bolzmann machines. By designing the first set of layers to be task dependent, and the deeper layers to be shared by both domains, useful representations can be learned that improve both retrieval and image classification. Moreover, tags can also be generated from this model. Gupta et al. [9] used prepositional relations and comparative adjectives to learn object detection models from weakly labeled data.

In [26], a corpus of webpages and non-aligned images are employed to learn a shared latent space using Kernel-CCA. Very promising improvements were obtained in the case of image annotation, however, performance of segmentation was rather poor. This is due to the difficulty of the problem, as a corpus of non-aligned news articles was employed. Li et al. [18] proposed a holistic generative model that reasons about classification, annotation (which classes are present in the scene) as well as segmentation. Tags are used to learn the holistic model with no other form of supervision. While these approaches are useful when labeled data is scarce, we would like to leverage additional supervision when available. Towards this goal, we utilize semantic labelings and rich textual descriptions of images to learn powerful holistic models. Importantly, our model is able to obtain very significant performance improvements in standard segmentation and detection tasks i.e., PASCAL VOC.

3. Automatic Text Extraction

In this section we show how to extract meaningful information from complex sentences for our image parsing task. We extract part of speech tags (POS) of all sentences using the Stanford POS Tagger for English language [28]. We parse syntactically the sentences and obtain a parse tree using the Stanford Parser with factored model [13]. Type dependencies were obtained using [5]. Highly efficient and

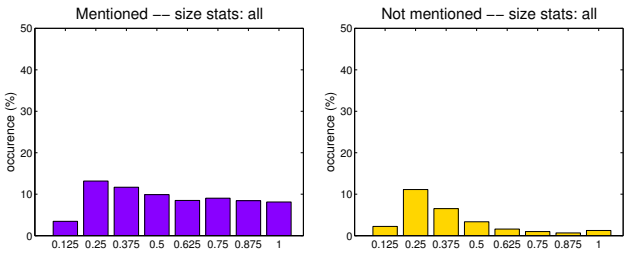


Figure 3. Size statistics when a class is mentioned or not

portable implementations of these algorithms are available online [1]. Fig. 2 shows an example of POS as well as parse trees obtained for our sentences. Given the POS, parse trees and type dependencies, we would like to extract whether an object class was mentioned as well as its cardinality (i.e., number of instances). Additionally, we are interested in extracting the relationships between the different objects, e.g., object A is near or on top of object B. We now discuss how to extract each type of information.

Presence of a Class: Text can be used to detect the presence/absence of a class. This can be done by extracting nouns from the POS, and matching those nouns to the object classes. In order to maximize the number of matched instances, we match nouns not only to the name of the class, but also to its synonyms and plural forms. This is important, as man, men and child are all synonyms of person.

Object Cardinality: The object cardinality can appear in the sentence in two different forms. First, it can be explicitly mentioned. For example in the sentence "two children playing on the grass", we only need to extract the word "two". This can be simply done by extracting the part of speech tag denoted CS, which encodes the cardinality. The second form is implicit, and arises when an explicit enumeration is not given. In this case, only a lower bound on the cardinality can be extracted. For this purpose, we parse the entire sentence from left to right and with each mention of a class noun we increase the count by 1 or 2, depending whether the word used is singular or plural. For example, we add 1 when we parse "child" and 2 when we parse "men". This processing is applied to each sentence individually.

Object Relations: We extract object relations by extracting prepositions and the objects they modify. Towards this goal, we first locate the prepositions of interest in the sentence (i.e., *near*, *in*, *on*, *in front of*). For each preposition we utilize the parse tree in order to locate the objects modified by the preposition. This allow us to create tuples of the form $(object_1, prep, object_2)$. Note that for a given preposition there could be more than one tuple of this form as well as more than one preposition in each sentence. For example in the sentence "two planes are parked near a car and a person", we can extract $(plane, near, car)$ and $(plane, near, person)$. To compute $object_2$ we search for NPs on the right side of the preposition by traversing the tree. We then return the nouns in the NP which are synonyms of our object classes. In the case of $object_1$, we move up the tree until we hit S or NP and return the nouns in the left child that are NPs which contain our object classes.

[1] http://nlp.stanford.edu/software/index.shtml

4. Holistic Scene Understanding

In this section we describe our approach to holistic scene understanding that uses text as well as image information in order to reason about multiple recognition tasks, i.e., segmentation, detection, scene classification. We formulate the problem as the one of inference in a CRF. The random field contains variables representing the class labels of image segments at two levels in a segmentation hierarchy (smaller and larger segments) as well as binary variables indicating the correctness of candidate object detections. In addition, binary variables encode the presence/absence of a class in the scene. Fig. 1 gives an overview of our model. Note that this type of structure has been successfully used for semantic parsing of still images [29, 16].

More formally, let $x_i \in \{1, \cdots, C\}$ be a random variable representing the class label of the i-th segment in the lower level of the hierarchy, while $y_j \in \{1, \cdots, C\}$ is a random variable associated with the class label of the j-th segment of the second level of the hierarchy. Following recent approaches [16, 17, 29], we represent the detections with a set of candidate bounding boxes. Let $b_l \in \{0, 1\}$ be a binary random variable associated with a candidate detection, taking value 0 when the detection is a false detection. Let $z_k \in \{0, 1\}$ be a random variable which takes value 1 if class k is present in the image.

We define our holistic CRF as

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{b}, \mathbf{s}) = \frac{1}{Z} \prod_{type} \prod_{\alpha} \psi_{\alpha}^{type}(\mathbf{a}_{\alpha}) \qquad (1)$$

where ψ_{α}^{type} encodes potential functions over sets of variables. Joint inference is then performed by computing the MAP estimate of the random field defined in Eq. 1.

In the following, we describe the different potentials that define our model. For clarity, we describe the potentials in the log domain, i.e., $w_{type}\phi_{\alpha}^{type} = \log(\psi_{\alpha})$. We employ a different weight for each type of potential, and share the weights across cliques. We learn the weights from training data using the structure prediction framework of [11].

We employ potentials which utilize the image I, text T, as well as statistics of the variables of interest. In addition, compatibility potentials relate the different tasks in the model (i.e., segmentation, detection and classification). We now describe the potentials we employed in detail.

4.1. Segmentation Potentials

Unary segmentation potential: We compute the unary potential for each region at segment and super-segment

	back.	aerop.	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	monitor	**averg.**
Textonboost (unary) [24]	77.8	14.1	3.4	0.7	11.3	3.3	25.5	30.9	10.3	0.7	13.2	10.8	5.2	15.1	31.8	41.0	0.0	3.7	2.4	17.1	33.7	16.8
Holistic Scene Understanding [29]	77.3	25.6	12.9	14.2	19.2	31.0	34.6	38.6	16.1	7.4	11.9	9.0	13.9	25.4	31.7	38.1	11.2	18.8	6.2	23.6	34.4	23.9
[29] num boxes from text	77.8	26.7	14.3	11.5	18.6	30.8	34.4	37.9	17.2	5.7	19.0	7.3	12.4	27.3	36.5	37.1	11.6	9.4	6.2	25.7	43.8	24.3
ours	76.9	31.3	29.7	37.3	27.7	29.5	52.1	40.0	38.0	6.6	55.9	25.2	33.2	38.2	44.3	42.5	15.2	32.0	20.2	40.7	48.4	36.4

Table 1. Comparison to the state-of-the-art that utilizes only image information in the UIUC sentence dataset. By leveraging text information our approach improves 12.5% AP. Note that this dataset contains only 600 PASCAL VOC 2008 images for training, and thus is significantly a more difficult task than recent VOC challenges which have up to 10K training images.

level by averaging the TextonBoost [16] pixel potentials inside each region. Thus, our segmentation potentials $\phi_s(x_i|I)$ and $\phi_s(y_j|I)$ depend only on the image.

Segment-SuperSegment compatibility: To encode long range dependencies as well as compatibility between the two levels of the hierarchy we use the P^n potentials of [14],

$$\phi_{i,j}(x_i, y_j) = \begin{cases} -\gamma & \text{if } x_i \neq y_j \\ 0 & \text{otherwise.} \end{cases}$$

Note that since we learn the weight associated with this potential, we are implicitly learning γ.

4.2. Class Presence Potentials

Class Presence from Text: We use two types of unary potentials, depending on whether a class was mentioned or not in the text. When a *class is mentioned*, we use the average cardinality (across all sentences) for each class

$$\phi^{class}_{ment}(z_i|T) = \begin{cases} \bar{C}ard(i) & \text{if } z_i = 1 \text{ and class } i \text{ mentioned} \\ 0 & \text{otherwise.} \end{cases}$$

When a *class is not mentioned* we simply use a bias

$$\phi^{class}_{notment}(z_i|T) = \begin{cases} 1 & \text{if } z_i = 0 \text{ and class } i \textbf{ not} \text{ mentioned} \\ 0 & \text{otherwise.} \end{cases}$$

We learn different weights for these two features for each class in order to learn the statistics of how often each class is "on" or "off" depending on whether it is mentioned.

Class Presence Statistics from Images: We employ the statistics of the training images in order to compute a unary potential over the presence and absence of each class z_i.

Class-Segment compatibility: This potential ensures that the classes that are inferred to be present in the scene are compatible with the classes that are chosen at the segment level. Thus

$$\phi_{j,k}(y_j, z_k) = \begin{cases} -\eta & \text{if } y_j = k \wedge z_k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

with η an arbitrarily large number, which is also learned.

4.3. Object Detection Potentials

We use DPM [8] to generate candidate object hypotheses. Each object hypothesis comes with a bounding box, a class, a score and mixture component id. In our model, we use the detections and reason about whether they are correct or not. In order to keep detections and segmentation coherent, each box b_ℓ is connected to the segments x_i it intersects with. Furthermore, each b_ℓ is connected to z to ensure coherence at the image level. We now describe the potentials.

Object Candidate Score: We employ DPM [8] as detector. It uses a class-specific threshold that accepts only the high scoring hypotheses. We reduce these thresholds to ensure that at least one box per class is available for each image. To keep the CRF model efficient, we upper bound the number of object hypotheses to be 3 per class. Thus, each images has at most 60 boxes (UIUC dataset has 20 classes). For each image, we use the boxes that pass the DPM thresholds, unless the object class is specifically mentioned in text. In this case, we add as many boxes as dictated by the extracted object cardinality $\bar{C}ard(cls)$. We utilize both text and images to compute the score for each detection. In particular, for each box we compute a feature vector composed of the original detector's score, the average cardinality for that class extracted from text, as well as object size relative to the image size. We use the latter since objects mentioned/not mentioned have different statistics (see Fig. 4). Since people usually tend to describe the salient objects, the non mentioned objects are usually small. We train a SVM classifier with a polynomial kernel with these features. Utilizing this classifier yields text-rescored object scores r_ℓ, which we use to re-rank the detections. We include the score in the model as a unary potential as follows:

$$\phi^{BBox}_{cls}(b_l|I, T) = \begin{cases} \sigma(r_l) & \text{if } b_l = 1 \wedge c_l = cls \\ 0 & \text{otherwise.} \end{cases}$$

Here c_l is the detector's class, and $\sigma(x) = 1/(1 + \exp(-1.5\,x))$ is a logistic function. We employ a different weight for each class in order to perform "context rescoring" within our holistic model.

Cardinality potential: We use a high-order potential on the b_ℓ variables to exploit the cardinality estimated from text. This does not slow down inference significantly as

	back.	aerop.	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	monitor	**averg.**
[29] num boxes from text	77.8	26.7	14.3	11.5	18.6	30.8	34.4	37.9	17.2	5.7	19.0	7.3	12.4	27.3	36.5	37.1	11.6	9.4	6.2	25.7	43.8	24.3
text rescored det.	77.9	28.1	12.8	31.0	32.3	32.3	44.3	42.1	27.4	6.4	26.0	22.1	24.1	29.5	37.3	40.8	11.4	21.6	16.5	26.8	45.1	30.3
+ card and scene	76.9	30.2	29.3	37.4	26.2	29.5	52.0	40.5	38.0	6.8	55.4	25.5	31.9	37.4	44.3	42.4	15.2	32.1	18.4	33.3	48.4	35.8
+ prep	76.9	31.3	29.7	37.3	27.7	29.5	52.1	40.0	38.0	6.6	55.9	25.2	33.2	38.2	44.3	42.5	15.2	32.0	20.2	40.7	48.4	36.4

Table 2. Performance gain when employing different amounts of text information. Our method is able to gradually increase performance.

	back.	aerop.	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	monitor	**averg.**
Oracle Z - noneg	76.8	36.7	28.3	34.8	21.9	30.9	56.1	47.6	36.8	10.0	58.2	28.8	33.4	54.8	42.6	41.8	15.1	28.1	16.3	35.7	48.7	37.3
Oracle Z - neg	76.8	31.2	28.2	34.7	21.7	31.1	56.0	50.3	36.7	10.5	57.4	29.5	34.4	55.1	42.5	41.9	16.9	32.2	18.8	35.7	49.2	37.7
ours	76.9	31.3	29.7	37.3	27.7	29.5	52.1	40.0	38.0	6.6	55.9	25.2	33.2	38.2	44.3	42.5	15.2	32.0	20.2	40.7	48.4	36.4

Table 3. Comparison to oracle Z (see text for details).

the high-order potential is over binary variables. In this potential we would like to encode that if the text refers to two cars being in the image, then our model should expect **at least** two cars to be on. Thus our potential should penalize all car box configurations that have cardinality smaller than the estimated cardinality. We thus define two potentials

$$\phi^{BBox}_{card-1}(\mathbf{b}^i|T) = \begin{cases} -\zeta_1 & \text{if } \bar{C}ard(i) \geq 1 \text{ and } \sum_j b^i_j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

and

$$\phi^{BBox}_{card-2}(\mathbf{b}^i|T) = \begin{cases} -\zeta_2 & \text{if } \bar{C}ard(i) \geq 2 \text{ and } \sum_j b^i_j = 1 \\ 0 & \text{otherwise.} \end{cases}$$

where $\mathbf{b}^i = \{b^i_1, b^i_2, \cdots\}$ refers to all detections of class i. We utilize two potentials to deal with noise in estimation, as the higher the cardinality the more noisy the estimate is.

Using prepositions: People tend to describe the objects in relation to each other, e.g., "the cat is on the sofa". This additional information should help boost certain box configurations that are spatially consistent with the relation. In order to exploit this fact, we extract prepositions from text and use them to score pairs of boxes. In particular, we extracted four prepositions (i.e., *near*, *in*, *on*, *in front of*) from text using the algorithm proposed in the previous section. For each relation, we get a triplet $rel = (cls_1, prep, cls_2)$. We train a classifier for each preposition that uses features defined over pairs of bounding boxes of the referred object classes, e.g., if the extracted relation was $\{person, in-front-of, car\}$, we take top 5 detections for person and car and form all possible pairs. For each pair, we extract the following spatial relation features: distance and signed distance between the closest left/right or top/bottom sides, amount of overlap between the boxes, and scores of the two boxes. We collect these feature vectors for each preposition ignoring the actual labels of the objects. We train for each preposition an SVM classifier with a polynomial kernel using these features. For each relation $rel = (cls_1, prep, cls_2)$ extracted from text, we form all possible pairs between the boxes for class cls_1 and class cls_2. We compute the new score for each box using the preposition classifier on the pairwise features as follows

$$\hat{r}_i = \max_{j,prep} score(r_{i,j,prep})$$

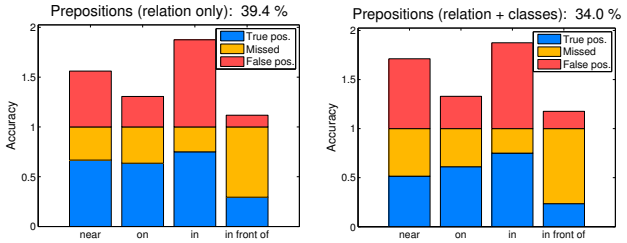


Figure 5. Text extraction accuracies: prepositions

where $r_{i,j,prep}$ is the output of the classifier for preposition $prep$ between boxes i and j. We do similarly for the box for cls_2. While order of classes (left or right to the preposition) might matters for certain prepositions (such as "on top of") in our experiments ignoring the position of the box yielded the best results. We then define unary potentials as

$$\phi_{prep}(b_\ell|T) = \begin{cases} \hat{r}_i & \text{if } b_l = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Class-detection compatibility: This term allows the bounding box to be active only when the class label of that object is also declared as present in the scene. Thus

$$\phi^{BClass}_{l,k}(b_l, z_k) = \begin{cases} -\alpha & \text{if } z_k = 0 \wedge c_l = k \wedge b_l = 1 \\ 0 & \text{otherwise.} \end{cases}$$

where α is a very large number estimated during learning.

Shape prior: The mixture components of a part-based model typically reflect the pose and shape of the object. Following [29], we exploit this by defining potentials $\phi^{sh}_{cls}(x_i, b_l|I)$ which utilize a shape prior for each component. Note that this feature only affects the segments inside the bounding box. We learn a different weight per-class, as the shape prior is more reliable for some classes than others.

4.4. Scene Potentials

Text Scene Potential: We first extract a vocabulary of words appearing in the full text corpus, resulting in 1793 words. We then train an SVM classifier with an RBF kernel over the bag-of-words (textual, not visual). As an additional feature, we use the average extracted cardinality for

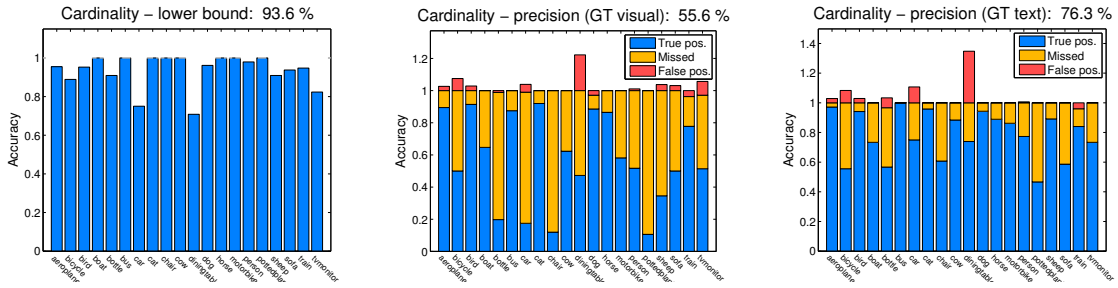


Figure 4. Text extraction accuracies: Cardinality

each class, which significantly boosted performance. We then define the unary potential for the scene node as

$$\phi^{Scene}(s = u|T) = \sigma(t_u)$$

where t_u denotes the classifier score for scene class u and σ is again the logistic function.

Scene-class compatibility: Following [29], we define a pairwise compatibility potential between the scene and the class labels as

$$\phi^{SC}(s, z_k) = \begin{cases} f_{s,z_k} & \text{if } z_k = 1 \wedge f_{s,z_k} > 0 \\ -\tau & \text{if } z_k = 1 \wedge f_{s,z_k} = 0 \\ 0 & \text{otherwise.} \end{cases}$$

where f_{s,z_k} represents the probability of occurrence of class z_k for scene type s, which is estimated empirically from the training images. This potential "boosts" classes that frequently co-occur with a scene type, and "suppresses" the classes that rarely co-occur with it.

4.5. Learning and Inference

Inference in this model is typically NP-hard. We rely on approximated algorithms based on LP relaxations. In particular, we employ the distributed convex belief propagation algorithm of [23] to compute marginals. For learning, we employ the primal-dual algorithm of [11] and optimize the log-loss. Following [29], we utilize a holistic loss which is the sum of the losses of the individual tasks. Intersection over union is used for detection, 0-1 loss for classification and pixel-wise labeling for segmentation.

5. Experimental Evaluation

To evaluate our method we use the UIUC dataset [7], which contains 1000 images taken from PASCAL VOC 2008. We use the segmentation annotations of [10], which has labels for most images. We labelled the missing images ourselves, and corrected some of the noisier MTurk annotations. Following [7], we use the first 600 images for training, and the remaining 400 for testing. Thus each class has ~ 50 training images. In addition, each image contains up to 5 sentences, 3 on average. We trained the DPM detectors using all VOC'10 trainval images excluding our test set.

Extracting information from text: We first evaluate the accuracy of the extracted information from text. We hand-annotated GT cardinality and prepositions for each sentence, which we call textGT. These annotations solely depend on text, and thus only the mentioned entities are contained in textGT, e.g., for man is walking with child, the textGT cardinality for person is two even though there may be more people in the image. This allows for a fair evaluation of our automatic text extraction. The top row of Fig. 4 shows the performance for our cardinality extraction technique with respect to: (a) Reliability – the number of times the image contains at least as many objects as the predicted cardinality, (b) visual information – the number of true, missed and extra predictions of the objects from text, compared to the visual GT, and (c) textGT – the number of true, missed and extra predictions of the object from text, compared to the textGT. Most missed cases belong to the most common and non-salient classes chair, car, potted-plant, bottle and bird. The plots show that the prediction of the lower bound on the number of objects is generally good (left), while the actual precision based on textGT is acceptable (right) and the precision with respect to visual GT is rather poor. In most cases the predicted cardinality is much lower than the actual number of objects in the image. This is likely due to the fact that people tend to describe the most interesting/salient object in the image and pay less attention to smaller and more common objects. However, despite the rather low prediction accuracy, our experiments show that cardinality is a very strong textual cue. Fig. 5 shows accuracy of preposition extraction from text.

Scene classifier: We used 15 scene types (i.e., dining area, room, furniture, potted-plant, cat, dog, city, motorbike, bicycle, field/farm, sheep, sky, airport, train, sea). Our text-based scene classifier achieved 76% classification accuracy. The best visual scene classifier achieved only 40%.

Holistic parsing using text: We next evaluate our holistic model for semantic segmentation. As evaluation measure, we employ the standard VOC IOU meassure. Our baselines consists of TextonBoost [25] (which is employed by our model to form segmentation unary potentials) as well as the holistic model of [29], which only employs visual information. As shown in Table 2, due to little training data the unary alone performs very poorly (17%). The holistic model of [29] achieves 23.9%. In contrast, by leveraging text, our approach performs very well, achieving 36.4%. Fig. 7 shows some examples of our inference.

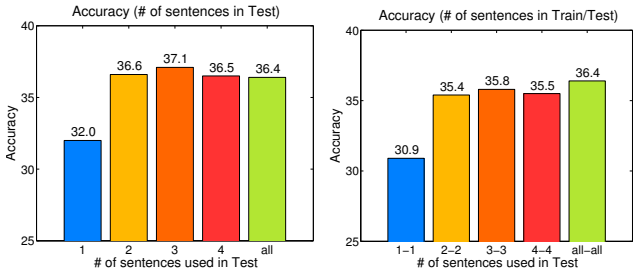


Figure 6. Segmentation accuracy as a function of the number of sentences used per image: **(a)** all sentences in training, and different number in test, **(b)** N in train and N test

.

Parsing with Oracle: Table 3 shows results obtained with our approach when using GT cardinality potentials instead of text cardinality potentials. Note that GT cardinality only affects the potential on z and potential on the number of detection boxes. The remaining potentials are kept the same. "GT noneg" denotes the experiment, where we encourage at least as many boxes as dictated by the cardinality to be on in the image. With "GT-neg" we denote the experiment where we also **suppress** the boxes for classes with card. 0. This means that for images where GT card. for a class is 0, we simply do not put any boxes of that class in the model.

Interestingly, by using GT instead of extracted cardinalities from text, we do not observe a significant boost in performance. This means that our holistic model is able to fully exploit the available information about the scene.

Model components: We next show how different components of our model influence its final performance. Rescoring the DPM detections via text gives significant improvement in AP, as shown in Table 4. We use these detections as candidate hypotheses in our model. Plugging the text-rescored detections directly into the model of [29] improves its performance to 30.3%. Adding the text-based cardinality and scene potentials, boosts the performance by another 5.5%, achieving 35.8%. Finally, using also the prepositions extracted from text, gives the final performance of 36.4%.

Amount of Text: We also experimented with the number of sentences per image. We tried two settings, **(a)** by using all available sentences per image in training, but different number in test, and **(b)** by varying also the number of training sentences. The segmentation results are shown in Fig. 5; As little as one sentence per image already boosts the performance to 32%. Since sentences can also not be directly related to the visual content of the image, having more of them increases performance. Two sentences already perform at the maximum. Similarly, re-scoring the detector in the same regime, shows that the detector achieves close to the best accuracy already with one sentence (Table 4). Fig. 5 shows results with different number of test sentences.

Time considerations: Our approach takes 1.8h for training the full dataset and 4.3s on average per image.

# train	1	2	3	4	all	all	all	all	all
# test	1	2	3	4	1	2	3	4	all
T-DPM	35.6	36.1	36.3	36.7	35.6	36.3	36.4	36.6	36.6
DPM	31.5								

Table 4. Results: Detector's AP (%) using text-based re-scoring for different number of sentences used per image in train and test.

6. Conclusions

In this paper, we tackled the problem of holistic scene understanding in scenarios where visual imagery is accompanied by text in the form of complex sentential descriptions or short image captions. We proposed a CRF model that employs visual and textual data to reason jointly about objects, segmentation and scene classification. We showed that by leveraging text, our model improves AP over vision-only state-of-the-art holistic models by 12.5%. As part of future work, we plan to exploit additional textual information (e.g., attributes) as well as leverage general text that accompanies images in webpages, e.g., Wikipedia.

References

[1] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. Video-in-sentences out. In *UAI*, 2012. 1, 2

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. In *JMLR*, 2003. 1, 2

[3] A. Berg, T. Berg, H. Daum, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *CVPR*, 2012. 1, 2

[4] D. Blei and M. Jordan. Modeling annotated data. In *ACM SIGIR*, 2003. 2

[5] M. de Marneffe, B. MacCartney, and C. Manning. Generating typed dependency parses from phrase structure parses. In *LREC*, 2006. 2

[6] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002. 1, 2

[7] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences for images. In *ECCV*, 2010. 1, 2, 6

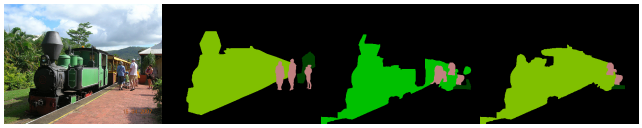
[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010. 1, 2, 4

[9] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008. 2

[10] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 6

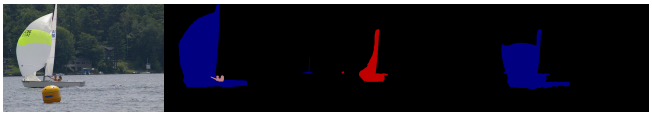
[11] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, 2010. 3, 6

sent 1: "A dog herding two sheep." **sent 2:** "A sheep dog and two sheep walking in a field." **sent 3:** "Black dog herding sheep in grassy field."

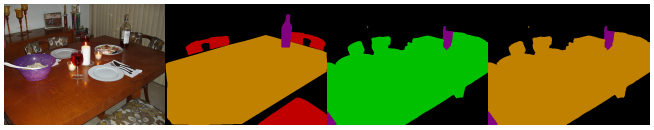


sent 1: "Passengers at a station waiting to board a train pulled by a green locomotive engine." **sent 2:** "Passengers loading onto a train with a green and black steam engine." **sent 3:** "Several people waiting to board the train."

sent 1: "Cattle in a snow-covered field." **sent 2:** "Cows grazing in a snow covered field." **sent 3:** "Five cows grazing in a snow covered field." **sent 4:** "Three black cows and one brown cow stand in a snowy field."



sent 1: "A yellow and white sail boat glides between the shore and a yellow buoy." **sent 2:** "Sail boat on water with two people riding inside." **sent 3:** "Small sailboat with spinnaker passing a buoy."



sent 1: "A table is set with wine and dishes for two people." **sent 2:** "A table set for two." **sent 3:** "A wooden table is set with candles, wine, and a purple plastic bowl."

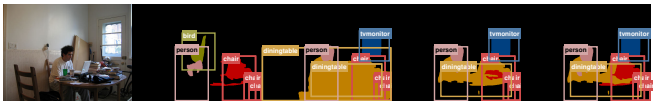
sent 1: "An old fashioned passenger bus with open windows." **sent 2:** "Bus with yellow flag sticking out window." **sent 3:** "The front of a red, blue, and yellow bus." **sent 4:** "The idle tourist bus awaits its passengers."

Figure 7. A few example images, sentences per image and the final segmentation.

sent 1: "Passengers at a station waiting to board a train pulled by a green locomotive engine." **sent 2:** "Passengers loading onto a train with a green and black steam engine." **sent 3:** "Several people waiting to board the train."

sent 1: "Black and white cows grazing in a pen.' **sent 2:** "The black and white cows pause in front of the gate." **sent 3:** "Two cows in a field grazing near a gate."

sent 1: "Two men on a plane, the closer one with a suspicious look on his face.'." **sent 2:** "A wide-eyed blonde man sits in an airplane next to an Asian man." **sent 3:** "Up close photo of man with short blonde hair on airplane."



sent 1: "Man using computer on a table." **sent 2:** "The man sitting at a messy table and using a laptop." **sent 3:** "Young man sitting at messy table staring at laptop."

Figure 8. Results as a function of the number of sentences employed.

[12] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, 2011. 1, 2

[13] D. Klein and C. Manning. Fast exact inference with a factored model for natural language parsing. In *NIPS'03*. 2

[14] P. Kohli, M. P. Kumar, and P. H. S. Torr. p^3 and beyond: Solving energies with higher order cliquess. In *CVPR'07*. 4

[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 1, 2

[16] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010. 3, 4

[17] D. C. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010. 3

[18] L. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding:classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009. 1, 2

[19] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 1, 2

[20] D. Putthividhy, H. Attias, and S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *CVPR*, 2010. 2

[21] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. In *CVPR07*. 1, 2

[22] K. Saenko and T. Darrel. Filtering abstract senses from image search results. In *NIPS*, 2009. 2

[23] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *CVPR*, 2011. 6

[24] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 4

[25] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *IJCV*, 81(1), 2009. 6

[26] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010. 1, 2

[27] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012. 1, 2

[28] K. Toutanova, D. Klein, and C. Manning. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*, 2003. 2

[29] Y. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 2, 3, 4, 5, 6, 7, 8