Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework

Li-Jia Li
Dept. of Computer Science
Princeton University, USA
jial@princeton.edu

Richard Socher
Dept. of Computer Science
Princeton University, USA
richard@socher.org

Li Fei-Fei
Dept. of Computer Science
Princeton University, USA
feifeili@cs.princeton.edu

Abstract

Given an image, we propose a hierarchical generative model that classifies the overall scene, recognizes and segments each object component, as well as annotates the image with a list of tags. To our knowledge, this is the first model that performs all three tasks in one coherent framework. For instance, a scene of a 'polo game' consists of several visual objects such as 'human', 'horse', 'grass', etc. In addition, it can be further annotated with a list of more abstract (e.g. 'dusk') or visually less salient (e.g. 'saddle') tags. Our generative model jointly explains images through a visual model and a textual model. Visually relevant objects are represented by regions and patches, while visually irrelevant textual annotations are influenced directly by the overall scene class. We propose a fully automatic learning framework that is able to learn robust scene models from noisy web data such as images and user tags from Flickr.com. We demonstrate the effectiveness of our framework by automatically classifying, annotating and segmenting images from eight classes depicting sport scenes. In all three tasks, our model significantly outperforms state-of-the-art algorithms.

1. Introduction

One of the most remarkable feats of the human visual system is how rapidly, accurately and comprehensively it can recognize and understand the complex visual world [7]. The various types of tasks related to understanding what we see in a visual scene is called 'visual recognition'. In computer vision, visual recognition has enjoyed some great success in recent years, particularly in single and/or isolated object categorization and localization. While recognizing isolated objects and object classes is a critical component of visual recognition, a lot more is needed to be done to reach a complete understanding of visual scenes. Take a picture of a polo game scene as an example. Often, within a single glance, humans are able to classify this image as a polo game (high-level scene classification), recognize different objects within the scene (annotation), and localize and delineate where the objects are in the scene (segmentation).

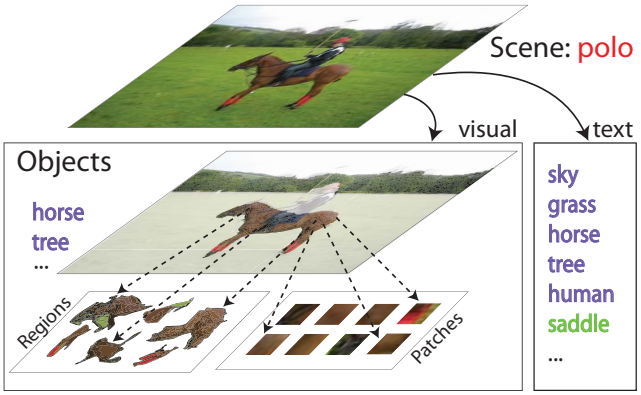

Figure 1. An example of what our model can understand given an unknown image. At the scene level, the image is classified as a 'polo' scene. A number of objects can be inferred and segmented by the visual information in the scene, hierarchically represented by object regions and feature patches. In addition, several tags can be inferred based on the scene class and the object correspondence.

No existing algorithm today can perform these tasks in a coherent framework. Towards this goal, we propose a unified framework to classify an image by recognizing, annotating and segmenting the objects within the image. The result of our algorithm is a generative model that encodes a hierarchy of semantic information contained in the scene (Fig.1). Three main motivations have guided our work. We highlight our contribution in achieving each of them in one unified framework.

Total scene understanding. Most of the earlier object and scene recognition work offers a single label to an image, e.g. an image of a panda, a car or a beach. Some go further in assigning a list of annotations without localizing where in the image each annotation belongs (e.g. [16]). A few concurrent segmentation and recognition approaches have suggested more detailed decomposition of an image into foreground object and background clutter. But all of them only apply to a single object or a single type of object (e.g. [15]). Our proposed model captures the co-occurrences of objects and high-level scene classes. Recognition becomes more accurate when different semantic components of an image are simultaneously recognized, allowing each component to

provide contextual constraints to facilitate the recognition of the others. In addition, both object recognition within a scene as well as scene classification can benefit from understanding the spatial extents of each semantic concept. *Our model can recognize and segment multiple objects as well as classify scenes in one coherent framework.*

Flexible and automatic learning. Learning scalability is a critical issue when considering practical applications of computer vision algorithms. For learning a single, isolated object, it is feasible to obtain labeled data. But as one wishes to understand complex scenes and their detailed object components, it becomes increasingly labor-intensive and impractical to obtain labeled data. Fortunately, the Internet offers a large amount of tagged images. *We propose a framework for automatic learning from Internet images and tags (i.e. flickr.com), hence offering a scalable approach with no additional human labor.*

Robust representation of the noisy, real-world data. While flickr images and tags provide a tremendous data resource, the caveat for exploiting such data is the large amount of noise in the user labels. The noisy nature of the labels is reflected in the highly uneven number and the quality of flickr tags: using a 'polo' image as an example, many tags do not have obvious visual correspondences (e.g. 'pakistan', 'adventure'); some tags can be incorrect (e.g. 'snow', 'mountain'); and visually salient tags are often missing (e.g. 'grass', 'human'). *Our generative model offers, for the first time, a principled representation to account for noise related to either erroneous or missing correspondences between visual concepts and textual annotations.*

In Sec.2, we describe the details of the model, its generative process and its properties. Sec.3 first illustrates how the parameters of the model are updated, and then provides an overview of the entire automatic learning framework. Sec.4 describes how classification, annotation and segmentation are performed given an unknown, unlabeled image. The subsequent Sec.5 will provide experimental results and model analysis on these three tasks. Lastly, Sec.6 compares our model to previous approaches.

2. The Hierarchical Generative Model

We propose a hierarchical generative model which aims to understand scene images, their objects and the associated noisy tags. The model shown in Fig.2 describes the scene of an image through two major components. In the visual component, a scene consists of objects that are in turn characterized by a collection of patches and several region features. The second component deals with noisy tags of the image by introducing a binary switch variable. This variable enables the model to decide whether a tag is visually represented by objects in the scene or whether it represents more visually irrelevant information of the scene. Therefore, the switch variable enables a principled joint modeling

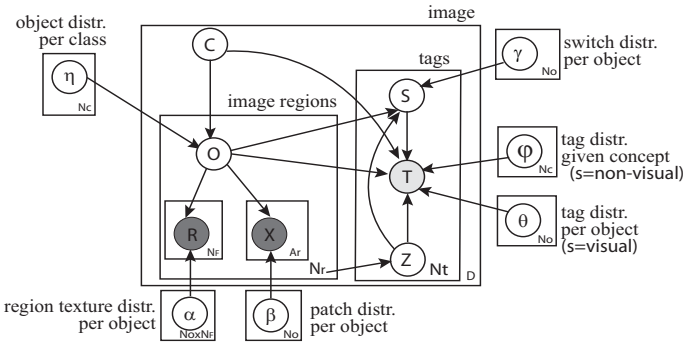

Figure 2. A graphical model representation of our generative model. Nodes represent random variables and edges indicate dependencies. The variable at the right lower corner of each box denotes the number of replications. The box indexed by N_r represents the visual information of the image, whereas the one indexed by N_t represents the textual information (i.e. tags). N_c, N_o, N_x, $N_{f_i}, i \in 1, 2, 3, 4$ denote the numbers of different scenes, objects, patches and regions for region feature type i respectively. Hyperparameters of the distributions are omitted for clarity.

of images and text and a coherent prediction of what tags are visually relevant. The hierarchical representation of image features, object regions, visually relevant and irrelevant tags, and overall scene provides both top-down and bottom-up contextual information to components of the model.

2.1. The Generative Process

In order to explain the generative process of our model mathematically, we first introduce the observable variables. Each image $d \in D$ is over-segmented into small coherent regions by using Felzenszwalb et al [9]. For each region, we extract $N_F = 4$ types of features, where $F = \{$shape, color, location, texture$\}$[1]. We further vector quantize region features into region codewords, denoted by the variable R in the model (see example of the representative regions for 'horse' in Fig.1). We use 100, 30, 50, 120 codewords for each feature type, respectively. Additionally, the set of patches X is obtained by dividing the image into blocks. Similarly, patches are represented as 500 codewords obtained by vector quantizing the SIFT [20] features extracted from them (see example of the representative patches for 'horse' in Fig.1). Noisy tags are represented by the variable T, which is observed in training. To generate an image and its corresponding annotations, a scene class C is sampled from a fixed uniform prior distribution. Given a scene, we are now ready to generate both the visual and textual components of the scene.

Generating the visual component. Given the scene class C, the probability of objects in such scenes is governed by a multinomial distribution. For each of the N_r

[1]We use the shape and location features described in [21]. Color features are simple histograms. Texture features are the average responses of filterbanks in each region.

image regions denoted by the left internal box in Fig.2, we first sample an object $O \sim \text{Mult}(\eta_c)$. Given the object O, we sample the image appearance:

1. For each $i \in F$, sample global appearance features: $R_i \sim \text{Mult}(\alpha_i|O)$, where there is a unique α_i for each object and each type of region feature.

2. Sample A_r many patches: $X \sim \text{Mult}(\beta|O)$.

Generating the tag component. At the same time, a region index Z is sampled from a uniform distribution. Z is used to account for the different numbers of tags and regions in this image, as suggested by [3]. As mentioned above, the switch variable S allows tags T to correspond to either visually relevant (i.e. the objects) or visually irrelevant (i.e. more abstract information) parts of the scene. This is formulated by allowing tags T to be drawn from either the distribution governed by object O, or the one controlled by scene class C. These ideas are summarized in the following generative procedure. For each of the N_t image tags:

1. Sample an index variable: $Z \sim \text{Unif}(N_r)$. Z is responsible for connecting an image region with a tag.

2. Sample the switch variable $S \sim \text{Binomial}(\gamma_{O_Z})$. S decides whether this tag is generated from the visually relevant object O or more visually irrelevant information related to the scene C. Fig.3 shows examples of switch probabilities for different objects.
   (a) If $S =$ non-visual: sample a tag $T \sim \text{Mult}(\varphi_c)$.
   (b) If $S =$ visual: sample a tag $T \sim \text{Mult}(\theta_{O_Z})$.

Putting the generative process together, the resulting joint distribution of scene class C, objects O, regions R, image patches X, annotation tags T, as well as all the latent variables becomes:

$$p(C, \mathbf{O}, \mathbf{R}, \mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{Z}|\eta, \alpha, \beta, \gamma, \theta, \varphi) = p(C) \cdot (\prod_{n=1}^{N_r} p(O_n|\eta, C)) \times \prod_{n=1}^{N_r} ((\prod_{i=1}^{N_F} p(R_{ni}|O_n, \alpha_i)) \cdot \prod_{r=1}^{A_r} p(X_{nr}|O_n, \beta)) \times \prod_{m=1}^{N_t} p(Z_m|N_r) p(S_m|O_{Z_m}, \gamma) p(T_m|O_{Z_m}, S_m, \theta, C, \varphi) \quad (1)$$

2.2. Properties of the Model

Our model is designed to perform three visual recognition tasks in one coherent framework: classification, annotation and segmentation. Eq.1 shows the joint probability of variables governing these three tasks. Later, in Eq.9, it will be clear how scene classification can directly influence the annotation and segmentation tasks.

Through the coupling of a scene C, its objects and their regions the model creates a hierarchical representation of an image. By modeling three layers jointly, they each improve the overall recognition accuracy. Each scene C defines a unique distribution $p(O|C)$ over objects. Additionally, the scene class C influences the distribution $p(T|C)$ over tags. This scene class influence serves as a top-down contextual facilitation of the object recognition and annotation tasks.

One unique feature of our algorithm is the concurrent segmentation, annotation and recognition, achieved by combining a textual and a visual model. Furthermore, our visual model goes beyond the 'bag-of-words' model by including global region features and patches inspired by [4].

Lastly, the model presents a principled approach to dealing with noisy tags. Fig.3 shows probability values of switch variable 'S' for different objects. When a tag is likely to be generated from a visually irrelevant, abstract source (e.g. 'wind'), its $p(S =$visual$)$ is low; whereas tags such as 'grass', 'horse' show high probability of being visually relevant.

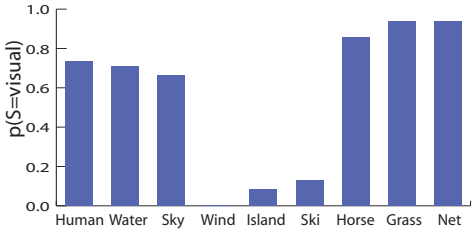

Figure 3. Probabilities of different objects. Words such as 'horse' or 'net' have higher probability because users tend to only tag them when the objects are really present, largely due to their clear visual relevance. On the contrary, words such as 'island' and 'wind' are usually related to the location or some other visually irrelevant concept, and usually not observable in a normal photograph.

3. Automatic Learning

We have described the model in detail and can now turn to learning its parameters. To this end, we derive a collapsed Gibbs sampling algorithm [24]. For each image and its tags, we sample the following latent variables: object O, switch variable S and index variable Z.

3.1. Learning via Collapsed Gibbs Sampling

Let O_{dn} denote the object for the nth region in the dth image, $\mathbf{R}_{dn}$ and $\mathbf{X}_{dn}$ represent the sets of its region features and patches. We define set $A_{dn} = \{j : Z_{dj} = n\}$ and $B_{dn} = \{j : Z_{dj} = n, S_{dj} = \text{visual}\}$. The switch variables related to A_{dn} is denoted as S_A, i.e., $S_A = \{S_{dj} : j \in A_{dn}\}$. The tags related to B_{dn} are represented as $T_B = \{T_{dj} : j \in B_{dn}\}$. $\overline{O}_{dn}$ represents all object assignments excluding O_{dn}. Similarly, we define the switch, index and tag variables S_{dm}, Z_{dm}, T_{dm}, $\overline{S}_{dm}$, $\overline{Z}_{dm}$ and $\overline{T}_{dm}$ for the mth tag in the dth image. $\overline{S}_A$ and $\overline{T}_B$ represent the corresponding assignments excluding A_{dn} and B_{dn} respectively. Following the Markov property of variable O, we analytically integrate out parameters $\eta, \alpha, \beta, \gamma, \varphi, \theta$. Then,

the posterior over the object O_{dn} can be described as:

$$p(O_{dn} = o|\overline{O}_{dn}, C_d, \mathbf{R}, \mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{Z}) \propto p(O_{dn} = o|\overline{O}_{dn}, C_d) \cdot p(\mathbf{R}_{dn}|\overline{\mathbf{R}}_{dn}, \mathbf{O}) \cdot p(\mathbf{X}_{dn}|\overline{\mathbf{X}}_{dn}, \mathbf{O}) \cdot p(\mathbf{Z}_A|N_r) \cdot p(S_A|\mathbf{O}, \mathbf{Z}, \overline{S}_A) \cdot p(T_B|\mathbf{O}, \mathbf{Z}, \mathbf{S}, \overline{T}_B) \quad (2)$$

Using standard Dirichlet integral formulation, we obtain the first element of this product:

$$p(O_{dn} = o|\overline{O}_{dn}, C_d = c) = \frac{n_{co,-dn} + \pi_o}{\sum_{o'} n_{co',-dn} + N_o \pi_o} \quad (3)$$

where π_o is the symmetric Dirichlet hyperparameter governing η. N_o is the total numbers of different objects. $n_{co,-dn}$ denotes the number of assignments of the object class o to scene class c, not including the current instance.

Similarly, the other counting variables $n_{of_i,-dn}, n_{ox,-dn}$, $n_{os,-dm}$ and $n_{ot,-dm}$ are also defined as the number of occurrences for f_i, x, s, t with o excluding the instances related to dn or dm. Given $n_{os} = \#(Z = z, O_z = o, S = s)$, $n_{os,-A_{dn}}$ indicates the frequency of s co-occurring with o excluding instances related to set A_{dn}. Given $n_{ot} = \#(Z = z, O_z = o, S = \text{visual}, T = t)$, $n_{ot,-B_{dn}}$ is the frequency of t co-occurring with o excluding instances related to set B_{dn}. $n_{ct,-dm}$ denotes the number of times tag t co-occurring with scene type c, excluding the current instance. Furthermore, N_{f_i}, N_x, N_t are the total numbers of different region features, patches and tags. The hyperparameters π_o, π_{f_i}, π_x, π_s, π_{ct}, π_{ot} are symmetric Dirichlet distributions governing $\eta, \alpha_i, \beta, \gamma, \varphi, \theta$.

S only has two possible values: $S = \text{visual}$ indicates a visually relevant object and $S = \text{non-visual}$ indicates a visually irrelevant object or scene information. Hence, $N_s = 2$.

The second and third part of Eq.3 become:

$$p(\mathbf{R}_{dn}|\overline{\mathbf{R}}_{dn}, \mathbf{O}) = \prod_{i=1}^{N_F} p(R_{dni} = f_i|\overline{\mathbf{R}}_{dni}, \mathbf{O}_{dn} = o, \overline{O}_{dn}) = \prod_{i=1}^{N_F} \frac{n_{of_i,-dn} + \pi_{f_i}}{\sum_{f_i'} n_{of_i',-dn} + N_{f_i}\pi_{f_i}} \quad (4)$$

$$p(\mathbf{X}_{dn}|\overline{\mathbf{X}}_{dn}, \mathbf{O}) = \frac{\Gamma(\sum_{x'} n_{ox',-dn} + N_x\pi_x)}{\Pi_{x'}\Gamma(n_{ox',-dn} + \pi_x)} \times \frac{\Pi_{x'}\Gamma(n_{ox'} + \pi_x)}{\Gamma(\sum_{x'} n_{ox'} + N_x\pi_x)} \quad (5)$$

$p(Z = z|N_r) = \frac{1}{N_r}$, hence $p(\mathbf{Z}_A|Nr)$ is constant. The part related to S is:

$$p(S_A|\mathbf{O}, \overline{S}_A, \mathbf{Z}) = \frac{\Gamma(\sum_{s'} n_{os',-A_{dn}} + N_s\pi_s)}{\Pi_{s'}\Gamma(n_{os',-A_{dn}} + \pi_s)} \frac{\Pi_{s'}\Gamma(n_{os'} + \pi_s)}{\Gamma(\sum_{s'} n_{os'} + N_s\pi_s)}. \quad (6)$$

The contribution of tags to object concept O_{dn} is:

$$p(T_B|\mathbf{O}, \overline{T}_B, \mathbf{Z}, \mathbf{S}) = \frac{\Gamma(\sum_{t'} n_{ot',-B_{dn}} + N_t\pi_{ot})}{\Pi_{t'}\Gamma(n_{ot',-B_{dn}} + \pi_{ot})} \frac{\Pi_{t'}\Gamma(n_{ot'} + \pi_{ot})}{\Gamma(\sum_{t'} n_{ot'} + N_t\pi_{ot})}. \quad (7)$$

Similarly, we derive the posterior over the switch variable S and index Z. Please see the technical report for details.

Up to this point, the update equations only need tags and images. However, there is no information of which tag T corresponds to which object O inside the image. Without such information it is possible to confuse tag-object relations, if both only occur together, e.g. 'water' and 'sailboat'. To prevent such a case, we introduce an automatic initialization system which provides a few labeled regions.

3.2. Automatic Initialization Scheme

In this section, we propose an initialization scheme which enables us to learn the model parameters with no human effort of labeling. The goal of the initialization stage is to provide a handful of relatively clean images in which some object regions are marked with their corresponding tags. During the learning process, these regions and tags provide seed information to the update equations.

In the preprocessing step, we use a lexicon to remove all tags that do not belong to the 'physical entity' group. Any lexicon dataset may be used for this purpose, we choose WordNet [22]. We also group all words in one WordNet synset (a group of synonyms) to one unique word, e.g. 'sailing boat' and 'catboat' are both transformed to 'sailboat'.

In the next step, we query flickr.com with the object names collected from the previous step to obtain initial training sets for each of the object classes. We then train the object model described in [4] and apply it to all scene images. A few object regions are collected from a handful of images for each object class. We now have a small number of partially annotated scene images and their still noisy tags, we select the best K of such images to seed the learning process described in Sec.3. This is done by ranking the images by favoring larger overlaps between the tags and annotated objects [2].

3.3. Learning Summary

Algorithm 1 summarizes the learning process. Furthermore, Fig.4 provides an example walk-through of one training image. After the pre-processing stage, some of the noisy tags are pruned out, but some visually relevant tags are still missing (e.g. tree, sky). The automatic initialization scheme then provides some segmented and annotated regions as seeds for the learning of the generative model. Using these seed images and additional unannotated images, the visual and textual components of the model are jointly trained. The effect of this joint learning is a robust model that can segment images more accurately (Fig.4(e)). Note that some visually irrelevant tags could also be recovered at the end of training (Fig.4(f), e.g. wind). This is attributed to

[2] $rank(O_d, T_d, P(T_d|C)) = \frac{\sum_{T_d \in O_d \cap T_d} P(T_d|C)}{\sum_{T_d \in O_d \cup T_d} P(T_d|C)}$, where T_d are the flickr tags of image d and $P(T_d|C)$ is the observed probability of tag T_d given the scene class C

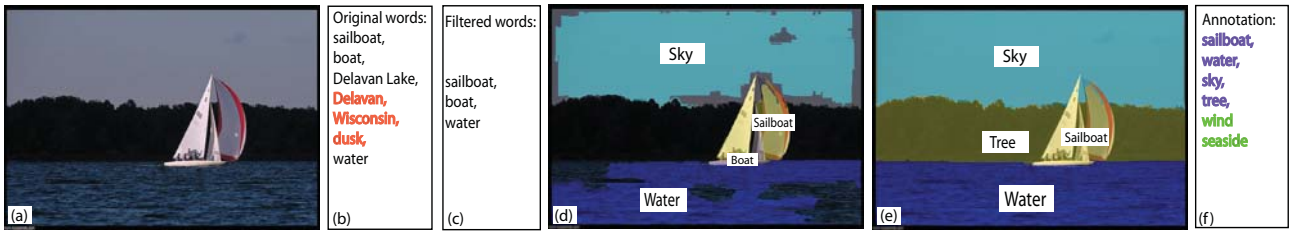

Figure 4. Walk-through of the learning process. **(a)** The original image. **(b)** The original tags from Flickr. Visually irrelevant tags are colored in red. **(c)** Output of Step 1 of Algorithm 1: Tags after the WordNet pruning. **(d)** Output of Step 2 of Algorithm 1: The image is partly annotated using the initialization scheme. Different object concepts are colored differently. Note that there is a background class in our initialization scheme, which is denoted in black in this figure. Since the criterion for being selected as an initial image is very conservative, the image annotations are clean but many regions are not annotated (missing tags). **(e):** Output of Step 3 of Algorithm 1: After training the hierarchical model, the image is completely and more precisely segmented. **(f):** Final annotation proposed by our approach. Blue tags are predicted by the visual component ($S = \text{visual}$). Green tags are generated from the top down scene information learned by the model ($S = \text{non-visual}$).

Algorithm 1 Automatic training framework

Step 1: Obtaining Candidate Tags Reduce the number of tags by keeping words that belong to the 'physical entity' group in WordNet. Group synonyms using WordNet synsets.

Step 2: Initialize Object Regions

Obtain initial object models. Apply the automatic learning method of [4] to learn an initial object model.

Annotate scene images. Apply the learned object model to annotate candidate object regions in each scene image.

Select initialization images. Select a small number of initialized images by a ranking metric described by Footnote2

Step 3: Automatic Learning. Treat the automatically selected top ranked images as 'supervised' data, add more flickr images and their tags to jointly train the model described in section 2.

the top-down influence of the scene class on the tags. Having learned this model, we can now turn to inference.

4. Inference: Classification, Annotation and Segmentation

Classification. The goal of classification is to estimate the most likely scene class for an image given an unknown image without any annotation tags. We use the visual component of the model (i.e. the region and patch appearances) to compute the probability of each scene class, by integrating out the latent object variable O:

$$p(C|\mathbf{R_d}, \mathbf{X_d}) = \frac{p(C, \mathbf{R_d}, \mathbf{X_d})}{p(\mathbf{R_d}, \mathbf{X_d})}$$
$$\propto \prod_{N_r} \sum_{O} p(\mathbf{R}|O)p(\mathbf{X}|O)p(O|C) \quad (8)$$

Finally, we choose $c = argmax_C p(C|\mathbf{R_d}, \mathbf{X_d})$.

Annotation. Given an unknown image, annotation tags are extracted from the segmentation results derived below.

Segmentation. Segmentation infers the exact pixel locations of each of the objects in the scene. By integrating out all the scene classes, we obtain:

$$p(O|\mathbf{R}, \mathbf{X}) = \sum_C p(O, C|\mathbf{R}, \mathbf{X}) \propto \sum_C p(O, C, \mathbf{R}, \mathbf{X})$$
$$= \sum_C p(O|C)p(\mathbf{R}|O)p(\mathbf{X}|O)p(C) \quad (9)$$

We observe that object segmentation is influenced both by the top-down force of scene class (first term in Eq.9) as well as the bottom-up force generated by the visual features (second and third terms in Eq.9).

5. Experiments and Results

We test our approach on 8 scene categories suggested in [18]: *badminton, bocce, croquet, polo, rock climbing, rowing, sailing, snowboarding*. By using these category names as keywords, we first automatically crawl the Flickr website to obtain 800 images and their tags for each category. 200 randomly selected images from each class are set aside as the testing images. After Step 1 of Algorithm 1, we obtain a vocabulary of 1256 unique tags. For segmentation experiments we consider the 30 most frequent words from this list. Note, however, that top down influence from the scene information still enables our model to be able to annotate images with tags from the full list of 1256 words. We offer more details about this dataset in the technical report.

Our hierarchical model can perform three tasks: image level classification, individual object annotation as well as pixel level segmentation. We now investigate performance of these tasks as well as the influence of parts of the model such as the switch variable on overall accuracy.

5.1. Scene Classification

The goal in this experiment is to classify an unknown image as one of the eight learned scene classes. We perform three experiments to analyze the different aspects of

our model and learning approach. All evaluations are done based on the 8-way classification results[3].

A. Comparison with different models. We compare the results of our model with three other approaches: (i) a baseline bag of words image classification model [8]; (ii) the region-based model used to initialize our initial object class models [4]; (iii) a modified Corr-LDA model based on [3] by adding a class variable on top of the mixing proportion parameter θ in the original model.

We provide the same list of tags generated by our system to our model and the modified model of [3]. Fig.5 summarizes the results. Our model consistently outperforms the other three approaches, whereas the region-based model [4] and the modified Corr-LDA model outperform a simple bag of words model. A comparison of the modified Corr-LDA model and our model underlines the effectiveness of our selective learning of visually relevant and irrelevant tags of the real-world data.

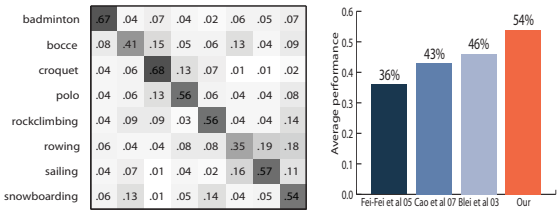

Figure 5. Comparison of classification results. **Left: Overall performance.** Confusion table for the 8-way scene classification. Rows represent the models for each scene while the columns represent the ground truth classes. The overall classification performance is 54%. **Right: Comparison with different models (Experiment A).** Performance of four methods. Percentage on each bar represents the average scene classification performance. $3rd$ bar is the modified Corr-LDA model [3].

B. Influence of unannotated data. To provide some insight into the learning process, we show in Fig.6-Left the classification performance curve as a function of the number of unlabeled images given to the model. In this experiment, the number of initialized images are fixed to 30. Performance gradually increases when more unlabeled images are included. This proves the effectiveness of unlabeled data in our learning framework.

C. Effect of noise in tags. In order to underline the robustness of our model to noisy training data, we present a set of experiments in which we dilute the original flickr tags with different percentages of noise by adding arbitrary words from the list of 1256 words during the training process. Fig.6-Right shows that while the algorithm of [3] decreases in accuracy when noise increases, our model is oblivious to even large percentages of noise. The robustness to noise is mostly attributed to the switch variable

[3]An 8-way classification result can be depicted by an 8×8 confusion table. By convention, we use the average of the diagonal entries of the table as the overall classification results of a particular model.

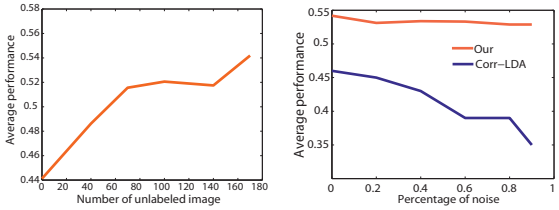

Figure 6. **Left: Influence of unannotated data (Experiment B).** Classification performance as a function of number of unannotated images. The y axis represents the average classification performance. The x axis represents the number of unlabeled images. It shows the unannotated images also contribute to the learning process of our model. **Right: Effect of noise in tags (Experiment C).** Performance of different models as a function of noise percentage in the tags. The y axis is average classification performance. The x axis represents the percentage of noisy tags. While the performance of corr-LDA decreases with the increase of percentage of noise, our model performs robustly by selectively learning the related tags.

'S', which correctly identifies most of the noisy tags, hence keeping the visual part of the model working properly even amidst a large amount of tagging noise.

5.2. Image annotation

Annotation tags are given through the results of segmentation. If there is a region of a certain object, we treat the name of this object as a tag.

D. Comparison to other annotation methods. In this experiment, we compare annotation results with two other state-of-the-art annotation methods – Alipr [16] and Corr-LDA [3]. We use precision-recall and F-measures to demonstrate the annotation results. Table 1 lists detailed annotation results for seven objects, as well as the overall scores from all object classes. Our annotation consistently and significantly outperforms the other two methods. This can be largely attributed to the selective learning of useful tags that can find a balance between bottom-up visual appearance cues and top-down scene class information.

	Alipr			Corr LDA			Our Model		
Object	Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
human	.83	.97	.89	.83	1.00	.91	.85	.98	**.91**
horse	–	–	–	.17	.91	.28	.17	.91	**.29**
grass	.42	.86	.56	.22	1.00	.35	.33	.86	**.48**
sky	.59	.33	.43	.55	.17	.26	.44	.92	**.59**
tree	.45	.38	.41	.25	.01	.03	.38	.93	**.54**
net	–	–	–	–	–	–	.27	.85	**.41**
sand	–	–	–	–	–	–	.24	.46	**.32**
Mean	**.15**	**.22**	**.16**	**.16**	**.40**	**.15**	**.28**	**.73**	**.34**

Table 1. Comparison of precision and recall values for annotation with Alipr, corr-LDA and our model. Detailed results are given for seven objects, but means are computed for all 30 object categories **(Experiment D)**.

5.3. Image segmentation

Our model not only classifies an image as a scene class, but also provides pixel level segmentation of the objects in

	Cao & Fei-Fei, 2007			Our Model		
Object	Prec	Rec	F	Prec	Rec	F
human	.35	.23	.28	.43	.47	**.45**
horse	.13	.49	.20	.27	.53	**.36**
grass	.62	.38	.47	.59	.50	**.54**
sky	.79	.44	.56	.74	.73	**.73**
tree	.40	.48	.44	.41	.59	**.48**
net	.04	.09	.05	.45	.26	**.33**
sand	.11	.32	.16	.29	.35	**.32**
Mean	**.22**	**.34**	**.28**	**.42**	**.46**	**.43**

Table 2. Results of segmentation on seven object categories and mean values for all 30 categories **(Experiment E)**.

the image without any such information given during training. We first compare quantitative results with another approach and then show a qualitative difference in example segmentations with and without the top down contextual influence provided by the scene class C.

E. Comparison to another segmentation method. In the image segmentation and annotation experiments, we train our model on 30 initialized images plus 170 unlabeled images. We test on 240 images where groundtruth is provided by human segmentation. Precision is computed by dividing the total area of correctly segmented pixels by the total area of detected pixels for each object. Recall is calculated by dividing the total area of correctly segmented pixels by the total area of true pixels of each object. We compare our segmentation results with the region-based model in [4]. [4] is used in the training of our initial object models. It is also one of the state-of-the-art concurrent object segmentation and recognition methods. Table 2 shows that our model significantly outperforms [4] in every object classes.

F. Influence of the scene class on annotation and segmentation. In this experiment, we examine the top-down, contextual influence of a scene in our model (Fig.7). We compare our full model to a damaged model in which the top down influence of the scene class is ignored. Our results underscore the effectiveness of the contextual facilitation by the top-down classification on the annotation and segmentation tasks.

6. Related Work

Our model is related to several research areas below.

Image understanding using contextual information. Semantically meaningful image understanding is a relatively recent topic in computer vision. A few earlier approaches have proposed interesting models for image understanding by object and scene, or multiple object recognition [31, 30, 14, 23, 12]. Also related are algorithms about object recognition in context, either through geometric constraints [13] or through semantic relations [26]. But none of these approaches has offered a rigorous probabilistic framework to perform simultaneous image classification, annotation and segmentation. Our earlier work [18] proposed a hierarchical model for event classification. But it only works in a fully supervised fashion, with cleanly segmented and annotated images.

Machine translation between words and images. Our work is also related to a family of models called 'pictures and words' models for images and annotations [1, 6, 5, 3][4] None of these models are capable of simultaneously performing image classification, annotation and segmentation. Furthermore, an important assumption of these models is that human annotations are provided in the training phase. None of them can robustly handle noisy tags.

Simultaneous object recognition and segmentation. A large body of work exists for simultaneous object recognition and segmentation [28, 15]. The object recognition and segmentation part of our model is related to a few recent unsupervised object categorization and segmentation papers [4, 27, 32]. But all of these models focus on delineating a single object class in images. No image level classification together with complete image annotation has be done.

Learning semantic visual models from Internet data. Finally, our approach is inspired by earlier approaches of learning from noisy internet data such as [2, 19, 10, 11, 29]. These approaches focus on single object classification to improve retrieval results from internet images. Our approach extends this to more exhaustive image understanding and explores the context correlation among objects.

7. Conclusion

In this paper, we have proposed a novel model to automatically classify, annotate and segment images of different scene classes. A hierarchical model is developed to unify the patch-level, object-level, and scene-level information. Our model is the first to provide a principled probabilistic treatment of noisy tags often seen in real-world data. Through an automatic training framework, we show that our model outperforms state-of-the-art methods in classifying, labeling and segmenting complex scene images. In the future, we will consider more sophisticated visual models to capture the geometry and appearance information of objects. We will also explore further the contextual relationships among objects within the scene.

Acknowledgement The authors would like to thank Chong Wang, Silvio Savarese, Bangpeng Yao, Hao Su, Barry Chai and anonymous reviewers for their helpful comments. Li Fei-Fei is funded by a Microsoft Research fellowship and a Google award.

References

[1] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3, 2003. 7

[2] T. Berg and D. Forsyth. Animals on the web. In *CVPR*, 2006. 7

[3] D. Blei and M. Jordan. Modeling annotated data. *SIGIR 03*. 3, 6, 7

[4] L. Cao and L. Fei-Fei. Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes. *ICCV*, 2007. 3, 4, 5, 6, 7

[4] Also related are video annotation and content-based image retrieval (e.g. [16, 17, 25]). But the limitations of these approaches are similar to the picture and words models.

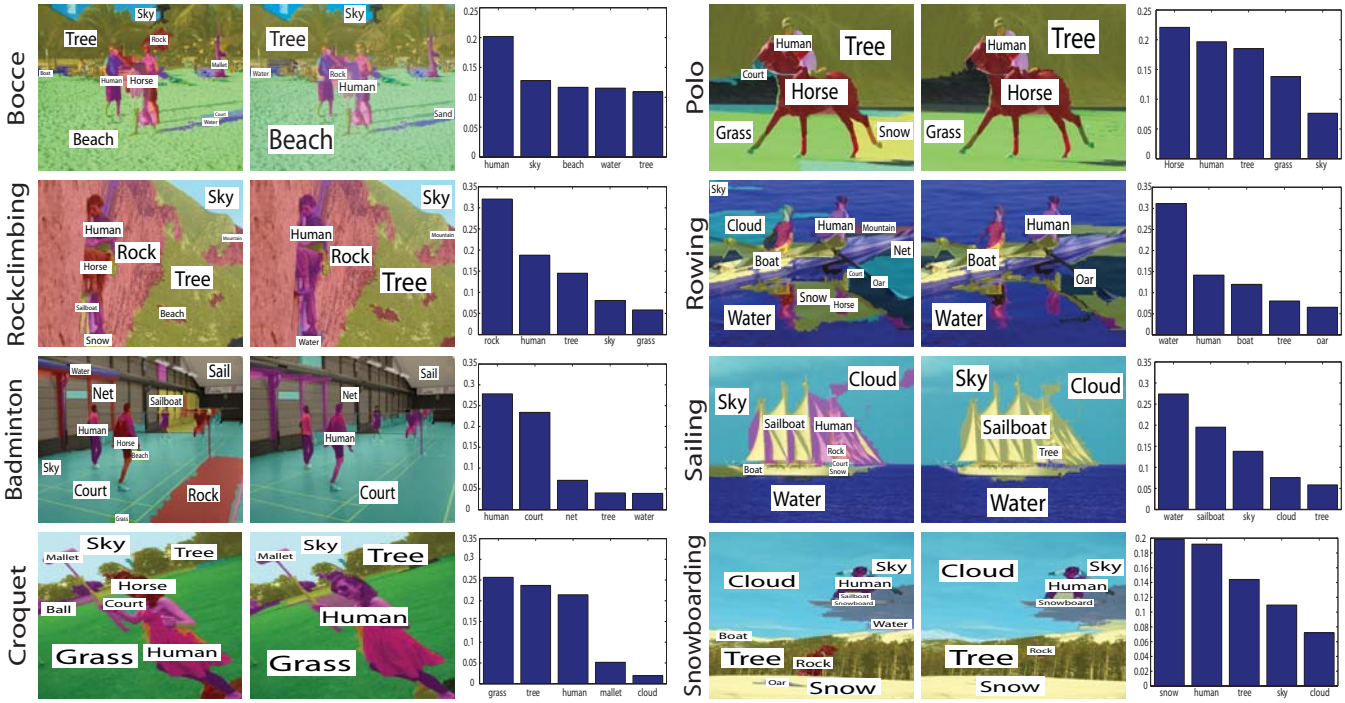

Figure 7. Comparison of object segmentation results with or without the top down scene class influence. Each triplet of images show results of one scene class **(Experiment F)**. The **left** image shows object segmentation result without the top down contextual information, i.e. by setting the probability distribution of object given scene class to a fixed uniform distribution. The **center** image shows object segmentation result by using the full model. We observe objects are more accurately recognized and delineated. The **right** image shows the probability of the 5 most likely objects per scene class. This probability encodes the top down contextual information.

[5] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. *ECCV*, 2004. 7

[6] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *European Conference on Computer Vision*, 2002. 7

[7] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we see in a glance of a scene? *Journal of Vision*, 2007. 1

[8] L. Fei-Fei and P. Perona. A Bayesian hierarchy model for learning natural scene categories. *CVPR*, 2005. 6

[9] P. Felzenszwalb and D. Huttenlocher. Efficient Graph-Based Image Segmentation. *IJCV*, 2004. 2

[10] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning Object Categories from Google Image Search. *ICCV 2005*, 2, 2005. 7

[11] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for Google images. In *ECCV*, 2004. 7

[12] A. Gupta and L. Davis. Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers. ECCV08. 7

[13] D. Hoiem, A. Efros, and M. Hebert. Putting Objects in Perspective. *CVPR*, 2006. 7

[14] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. *CVPR*, 2006. 7

[15] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj cut. In *Computer Vision and Pattern Recognition*, 2005. 1, 7

[16] J. Li and J. Wang. Automatic Linguistic Indexing of Pictures by a statistical modeling approach. *PAMI*, 2003. 1, 6, 7

[17] J. Li and J. Wang. Real-time computerized annotation of pictures. In *In Proc. ACM Multimedia*, 2006. 7

[18] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Proc. ICCV*, 2007. 5, 7

[19] L.-J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. In *Proc. CVPR*, 2007. 7

[20] D. Lowe. Object recognition from local scale-invariant features. In *Proc. International Conference on Computer Vision*, 1999. 2

[21] T. Malisiewicz and A. A. Efros. Recognition by association via learning per-exemplar distances. In *CVPR*, 2008. 2

[22] G. Miller. WordNet: A Lexical Database for English. *COMMUNICATIONS OF THE ACM*, 1995. 4

[23] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees:a graphical model relating features, objects and scenes. In *NIPS (Neural Info. Processing Systems)*, 2004. 7

[24] R. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS*, 9(2):249–265, 2000. 3

[25] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, and H. Zhang. Correlative multi-label video annotation. In *ACM Multimedia*, 2007. 7

[26] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 7

[27] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. *Proc. CVPR*, 2006. 7

[28] E. Sali and S. Ullman. Combining class-specific fragments for object classification. In *BMVC*, 1999. 7

[29] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV 2007*. 7

[30] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005. 7

[31] Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image Parsing: Unifying Segmentation, Detection, and Recognition. *IJCV*, 2005. 7

[32] X. Wang and E. Grimson. Spatial Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems*, 2007. 7