Weakly Supervised Learning for Attribute Localization in Outdoor Scenes

Shuo Wang[1,2], Jungseock Joo[2], Yizhou Wang[1], and Song-Chun Zhu[2]

[1]Nat'l Engineering Lab for Video Technology, Key Lab. of Machine Perception (MoE), Sch'l of EECS,
Peking University, Beijing, 100871, China
{shuowang, Yizhou.Wang}@pku.edu.cn

[2]Department of Statistics, University of California, Los Angeles (UCLA), USA
sczhu@stat.ucla.edu, joo@cs.ucla.edu

Abstract

In this paper, we propose a weakly supervised method for simultaneously learning scene parts and attributes from a collection of images associated with attributes in text, where the precise localization of the each attribute left unknown. Our method includes three aspects. (i) Compositional scene configuration. *We learn the spatial layouts of the scene by Hierarchical Space Tiling (HST) representation, which can generate an excessive number of scene configurations through the hierarchical composition of a relatively small number of parts.* (ii) Attribute association. *The scene attributes contain nouns and adjectives corresponding to the objects and their appearance descriptions respectively. We assign the nouns to the nodes (parts) in HST using non-maximum suppression of their correlation, then train an appearance model for each noun+adjective attribute pair.* (iii) Joint inference and learning. *For an image, we compute the most probable parse tree with the attributes as an instantiation of the HST by dynamic programming. Then update the HST and attribute association based on the inferred parse trees. We evaluate the proposed method by (i) showing the improvement of attribute recognition accuracy; and (ii) comparing the average precision of localizing attributes to the scene parts.*

1. Introduction

In the past decade, researchers have made significant progress in scene categorization [1, 16, 18]. Most of the popular methods first exact features, such as scene gist [1], spatial pyramid [16] and Tangram [8], then feed to SVM classifiers. In contrast to basic level scene categorization, natural scenes often contain semantic details that might be attributed to more than one category. Thus the interest in studying the scene attributes [6, 7] has been growing. A typical recent work is by Patterson and Hays [7] which identified 102 scene attributes through human perception experiments and trained 102 independent classifiers. Such methods obtained interesting results and are potentially useful for image retrieval, however, they have some obvious limitations: the attributes are not associated with the specific image regions (called "scene parts" in the following), and not explicitly linked to the appearance models of the parts.

In this paper, we propose a weakly supervised method to study the scene configuration and attribute localization. As shown in Fig. 1, our approach begins with a collection of images with attributes in text (Fig.1(a)). The training images are labeled with the presence of several attributes, with the precise localization of the attributes left unknown. Our method includes three aspects as below.

(i) Hierarchical scene configuration and part learning. A typical scene category, *e.g.* countryside or city street, contains a huge number of configurations with objects (buildings, road *etc.*) and regions (sky, field *etc.*) in different layouts. To learn the meaningful hierarchy and scene parts, we utilize the Hierarchical Space Tiling (HST) [17] to represent the scenes. As shown in the top row of Fig.1(b), the HST quantizes the huge space of scene configurations by a stochastic And-Or Tree (AOT) representation where an And-node represents a way of decomposing the node, an Or-node represents alternative decompositions, and the terminal nodes are primitive rectangles corresponding to the scene parts. Through a learning-by-parsing strategy, we can learn the HST/AOT model and a scene part dictionary, in which each scene part corresponds to a meaningful region in the scenes such as sky, building, road, field.

(ii) Attribute association. Scene attributes, defined by the text descriptions, consist of the nouns (*e.g.* field, sky) and adjectives (*e.g.* green, cloudy), corresponding to the ob-
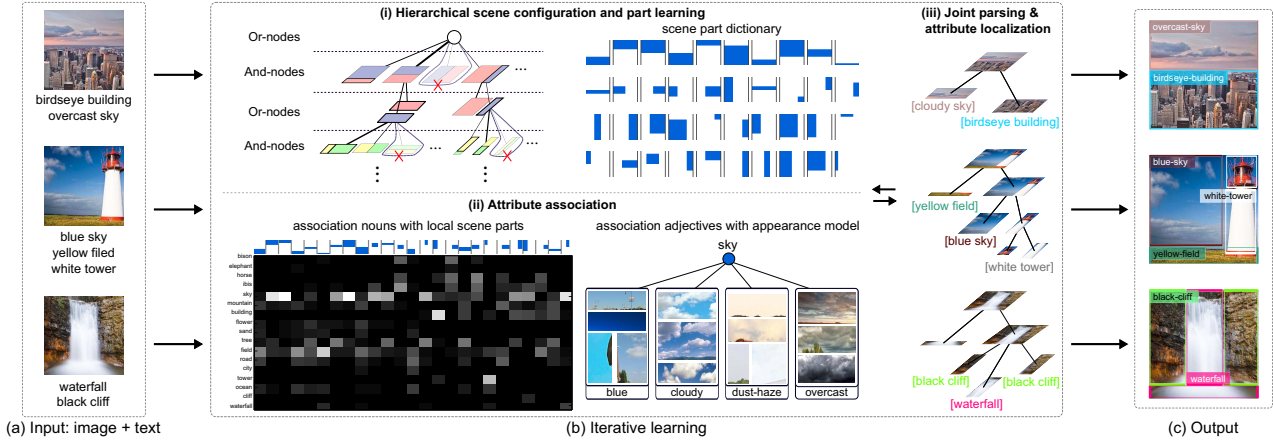
IEEE computer society

Figure 1. Flowchart of our method. (a) Input images and texts. (b) Iterative learning process including the learning of scene configuration and attribute association, and the joint inference (text in square brackets denotes the inferred attributes). (c) Output attribute localization.

jects/regions and their appearance respectively. The nouns are assigned to the learned scene part dictionary according to an association matrix as shown in the bottom left of Fig.1(b). The association matrix measures the probability of a noun and a scene parts appearing simultaneously in the training set and it can be achieve by a non-maximum suppression. Each noun has a mixture of appearance models corresponding to the adjectives, *e.g.* the sky may be blue, cloudy, dust-hazed or overcast (bottom right of Fig.1(b)).

(iii) Joint inference and learning. Given an image, we jointly infer the optimal parse tree and localize the semantic attributes to the scene parts by dynamic programming (right panel in Fig.1(b)). Then based on the inferred parse trees, we re-estimate the HST/AOT model and attribute association matrix. Thus, we integrate the parsing and attribute localization under an uniform framework.

We evaluate the proposed method by showing: (i) The semantic attributes are properly associated with the local scene parts. (ii) Compared with traditional classification algorithms, our method achieves better attribute recognition performance. (iii) We improve the precision of attribute localization against a baseline sliding window method [10].

2. Related Work

Scene models   For scene classification, there are four typical representations. (i) *Bag-of-Words (BoW) representation* [11] treats a scene as a collection of visual words and ignores the spatial information. (ii) *Grid structure representation*, such as spatial pyramid matching [16], implicitly adopt squares as elements in different sizes and locations and divide the images into grids. (iii) *Non-parametric representation*, such as label transfer [4], remembers all the observed images and interprets the new data through nearest neighbor search. All these representations miss the hierarchical reconfigurable structures. (iv) The most related work is the Hierarchical Space Tiling (HST) [17] which in-

troduced a scene hierarchy by the And-Or Tree (AOT) and proposed a structure learning method to learn a scene part dictionary and compact HST model. However, it relies on the label maps as training samples. We extend [17] to take raw images with text as input and associate scene attributes to the learned scene part dictionary.

Scene attributes   Beyond recognizing an individual scene category, visual attributes are demonstrated as valuable semantic cues in various problems such as generating descriptions of unfamiliar objects [6]. Patterson and Hays [7] proposed an attribute based scene representation containing 102 binary attributes to describe the intra-class scene variations (*e.g.* a canyon might have water or it might not) and the inter-class scene relationships (*e.g.* both a canyon and a beach could have water). Beside the binary attributes, Parikh and Grauman [5] introduced the relative attributes, *e.g.* more natural or less man-made, to provide a semantically rich way describing and comparing scenes. These attributes were learned and inferred at the image level without localization. In contrast, we jointly parse the images into spatial configurations and localize the attributes, which allows us to provide more accurate and detailed descriptions.

Attributes localization   In learning the relationships between the attributes and specific image regions, we relate to the recent work on object detection and localization. The two communities of object localization include sliding window based methods and Multiple Instance Learning (MIL). (i) The sliding window methods [10] operate by evaluating a classifier function at many different sub-windows of the image and then predicting the object presence in sub-windows with high-score. (ii) Multiple Instance Learning (MIL) based algorithms [2, 3] view images as bags of segments. Then MIL trains a binary classifier to predict the class of segments, under the assumption that each positive training image contains at least one true-positive seg-
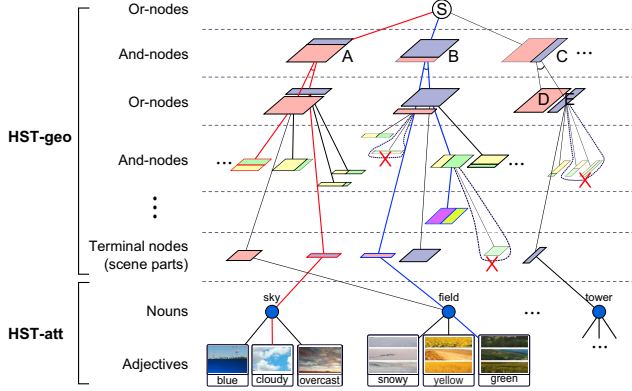
Figure 2. HST scene representation which consists of HST-geo and HST-att.

ment. However, these approaches incur the problem faced by the unreliable segmentation. The above methods can localize an individual object at a time, while our aim is parsing the images into multiple objects/attributes simultaneously. Moreover, by considering the global compatibility, our method will not confuse the objects with similar appearance (*e.g.* "blue ocean" and "blue sky") as the above methods did.

3. Representation: HST/AOT

As shown in Fig.2, we extend the original HST[17] to two parts: (i) *HST-geo* which models the geometric arrangements of the scenes, *i.e.* scene configurations, and (ii) *HST-att* which models the appearance types of the scene attributes and the correlations between the scene parts and attributes.

For HST-geo, there are three types of nodes: *Or-nodes* V^{OR}, *And-nodes* V^{AND}, and *Terminal nodes* V^T.

The Or-nodes V^{OR}, correspond to the grammar rules like $r^{OR} : S \rightarrow A|B|C$, acting as "switches" between the possible compositions. The branching probabilities $p(A|S)$, $p(B|S)$, $p(C|S)$ indicate the preference for each composition and can be learned from the scene images in Section.4.1.

The And-nodes V^{AND}, correspond to the grammar rules like $r^{AND} : C \rightarrow D \cdot E$, representing a fixed decomposition from a node C into lower-level parts D and E. For simplicity, we only divide the rectangular parts in horizontal and vertical ways.

The terminal nodes V^T, form a scene part dictionary $\Delta = V^T$. At the bottom of the hierarchy, an image lattice is divided into a $n \times n$ grid, and each cell is seen as an atomic shape element of the dictionary. A number of the atomic elements compose the higher-level terminal nodes at different scales, locations and shapes. To avoid the combination explosion, only regular shapes *i.e.*, squares, rectangles are allowed.

Beyond HST-geo, we combine the scene attributes to

represent both the geometry and semantics of the scenes. Scene attributes come from the text descriptions of training images which contain several noun+adjective phrases. The nouns correspond to the objects in the scenes and the adjectives correspond to the appearance. We model the HST-att as a two level AOT. Each noun, acting as an appearance-Or node, has a mixture of adjectives. And each terminal node in HST-geo can link to a noun and further an adjective attribute by an association matrix. The association matrix can be learned in Section.4.2.

The HST is naturally recursive, starting from a root which is an Or-node, generating the alternating levels of And-nodes and Or-nodes, and stopping at the terminal nodes with a specific appearance type (noun+adjective). The And-Or structure defines a space of possible parse trees and embodies probabilistic context free grammar (PCFG) [15]. By selecting the branches at Or-nodes, a parse tree pt is derived, *e.g.* the red and blue paths in Fig.2 represents two parse trees as instances of the HST. When parse trees collapse to the image lattice, they produce configurations.

The initial HST is excessive and generates a combinatorial number of parse trees. In the learning process, we maximize the likelihood subject to a model complexity and prune out the branches with zero or low probability to obtain a compact HST and the scene part dictionary.

4. Learning

4.1. Learning for the HST-geo

We define the HST-geo as a 4-tupe

$$HST\text{-}geo = (S, V^N, V^T; \Theta) \tag{1}$$

where S is a start symbol at root. $V^N = V^{AND} \cup N^{OR}$ is a set of non-terminal nodes including the And-nodes and Or-nodes. V^T is a set of terminal nodes forming the scene part dictionary $\Delta = V^T$. Let v index the nodes; $Ch(v)$ denote the child node set of v. The parameters Θ are the branching probabilities of each branch at the Or-nodes $\Theta = \{\theta(v \rightarrow v_i); v \in V^{OR}, v_i \in Ch(v)\}$.

Given a set of training images $\mathbf{I} = \{I_m, m = 1...M\}$, in order to avoid the false compositions, *e.g.* sky and ocean may be grouped wrongly into one region due to their similar appearance, we first segment the images in multi-scale in a coarse-to-fine manner so that we can focus the learning on the label maps and thus separate the geometric configurations from appearance. Let $\mathbf{C} = \{C_m, m = 1...M\}$ denotes the multi-scale segmentation. $C_m = \{C_m^k\}$ includes $|k|$ segmented layers. For each image, we adopt [13] to obtain the multi-scale segmentation by tuning $k \in 300, 400, ..., 5000$, where k is a variable controlling the granularity of the segmentation (Fig.3). Then we select six segmented layers of significant difference by comparing the
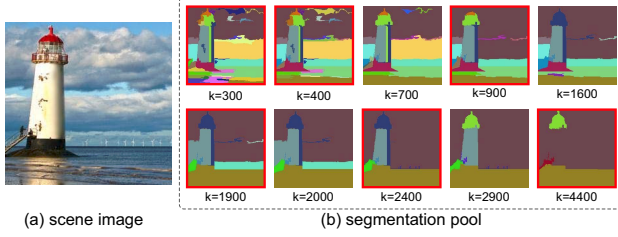
Figure 3. Multi-scale segmentation. (a) Input image. (b) Segmentations in different layers. The segmentations in the red frames form a multi-scale segmentation C.

adjacent layers in pixels (red frames in Fig.3(b)), and compose a multi-scale segmentation set $C = \{C^k, |k| = 6\}$.

The learning requires us to estimate the branching probabilities Θ and scene part dictionary Δ by maximizing a log-likelihood.

$$(\Theta, \Delta)^* = \arg\max_{\Theta, \Delta} \log p(\mathbf{I}; \Theta, \Delta) \quad (2)$$

$$\propto \arg\max_{\Theta, \Delta} \sum_{m=1}^{M} \log \sum_{pt_m, k} p(C_m^k, pt_m; \Theta, \Delta) p(I_m | C_m^k)$$

where $\log p(I|C^k) = -\sum_{c \in \{R,G,B\}} \sum_{r \in C^k} \sigma_c^2(I(r))$, c is a color channel, r is a segment in C^k, $\sigma()$ returns the standard divination of the pixel intensities in a segment. This term measures the pixel intensity homogeneity of the segments. $p(C^k, pt; \Theta, \Delta)$ is the joint probability with Θ and Δ being the parameters to be learned.

$$\log p(C^k, pt; \Theta, \Delta) \propto -E(C^k, pt; \Theta, \Delta) \quad (3)$$

$$= -\sum_{v \in V_{pt}^{OR}, v_i \in Ch(v)} E^{OR}(v_i|v) - \lambda \sum_{v \in V_{pt}^{T}} E^T(C_v^k|v)$$

where V_{pt}^{OR}, V_{pt}^{T} denote the Or-nodes and terminal nodes in the pt, and λ is the parameter to balance the two terms ($\lambda = 0.25$ in this paper). C_v^k denotes the segmented patch covered by the terminal node v.

The energy for an Or-node is defined on its branching probability, which favors the sub-structures that often make a larger part. *i.e.,*

$$E^{OR}(v_i|v) = -\ln \theta(v \to v_i) \quad (4)$$

The energy for a terminal node is defined as

$$E^T(C_v^k|v) = -\ln \frac{1}{|C_v^k|} \sum_{i \in C_v^k} \mathbb{1}[l_i^k = l_v^k] + \ln \frac{k}{|C_v^k|} \quad (5)$$

where $\mathbb{1}[\cdot]$ is the indicator function. In the k-th layer, l_i^k is the segmentation label of pixel i and l_v^k is the dominant label of the terminal node v. The first term measures the homogeneity of the terminal nodes in terms of segmentation labels and the second term penalizes large k.

Iterative learning of the HST-geo To maximize the Eq.2, we adopt an iteratively learning-by-parsing strategy

including: (i) inferring the optimal parse tree pt by dynamic programming (optimize Eq.3); and (ii) estimating the parameters Θ by a maximum likelihood estimator (MLE). After it converges, those branches whose probabilities are below a certain threshold (say 0.01) are pruned. Then we collect the terminal nodes from all the parse trees to form a scene part dictionary Δ. (see more details in [17]).

Terminal node local adjustment Although the scenes from one category share similar spatial layouts, there are still considerable variations/deformations in their configurations. Hence, the terminal nodes are allowed to be locally adjustable to fit the scene region boundaries. We introduce the perturbations in location, scale and orientation denoted as $\delta(x) = [\pm 8, \pm 16]$, $\delta(s) = [1 \pm \frac{1}{32}, 1 \pm \frac{1}{16}]$ and $\delta(a) = [\pm \frac{\pi}{48}, \pm \frac{\pi}{24}]$, respectively. Thus the total number of node activities is 12 in addition to the original one.

4.2. Learning for the HST-att

The text descriptions usually contain noun+adjective phrases: The nouns indicate objects/regions inside a scene (*e.g.* sky, field); and the adjectives describe their appearance (*e.g.* overcast, green). Let $\mathcal{A} = \{\mathcal{A}^n, \mathcal{A}^{adj}\}$ denote the attribute set, where $\mathcal{A}^n$ is the noun attribute set and $\mathcal{A}^{adj}$ is the adjective attribute set.

We explore the relationship between a noun $a \in \mathcal{A}^n$ and a scene part $v \in \Delta$ by an association matrix:

$$\Phi : \mathcal{A}^n \times \Delta \mapsto [0,1], \text{s.t.} \sum_{a \in \mathcal{A}^n} \Phi(a, v) = 1, \forall v \in \Delta \quad (6)$$

where the entries of the rows in Φ are the noun attributes and the columns are the scene parts, and we normalize each columns to be one.

After learning the HST-geo in Section.4.1, each training image has an optimal parse tree pt. Because the attributes are annotated at the image level rather than the precise image regions, we initialize Φ by counting all the combinations of the nouns and the terminal nodes in pt:

$$\Phi(a, v) = \sum_{m=1}^{M} \mathbb{1}[a \in A_m^n] \cdot \mathbb{1}[v \in pt_m] \cdot \phi_m(a, v) \quad (7)$$

where $A_m^n \subseteq \mathcal{A}^n$ is the noun attribute set for an image, and $\phi_m(a, v)$ denotes its association probability initialized by $\phi_m(a, v) = 1$.

We pursue Φ by a greedy non-maximum suppression. The algorithm first selects an (a, v) pair which receives the highest association probability: $(a^*, v^*) = \arg\max_{(a,v)} \Phi$, and find the image set $\tilde{I} \subseteq I$ having (a^*, v^*), *i.e.* $\tilde{I} = \{I_m; a^* \in A_m^n, v^* \in V_{pt_m}^T\}$. Then (i) suppress the association between the selected attribute with other terminal nodes except v^*: $\phi_m(a^*, v) = s \times \phi_m(a^*, v); v \in V_{pt_m}^T \setminus v^*, I_m \in \tilde{I}$, where $s = 0.3$ is the suppression parameter; (ii) suppress the association between the selected node with other noun attributes: $\phi_m(a, v^*) = s \times \phi_m(a, v^*); a \in A_m^n \setminus a^*, I \in \tilde{I}$;
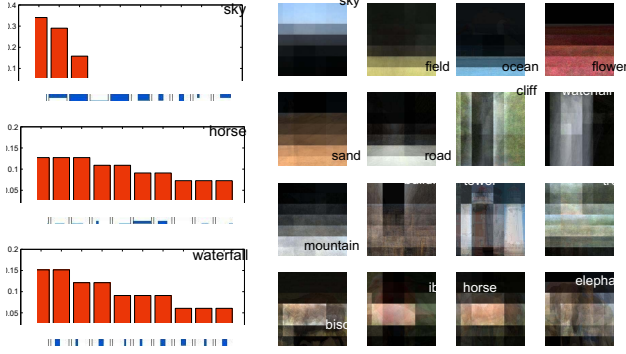
Figure 4. The association of noun attributes and the scene parts.

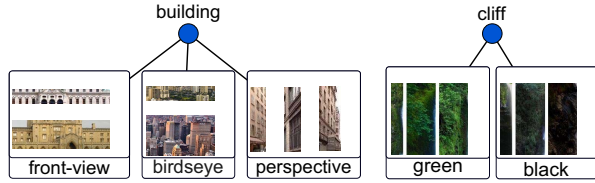Figure 5. The adjective clusters belonging to the noun attributes.

(iii) update Φ by Eq.7. Repeatedly find the next maximum (a, v) pair and do non-maximum suppression until no more (a, v) pair can be selected. Finally, normalize each columns in Φ to be one.

Fig.4 (left) shows the association of noun attributes and scene parts, where the horizontal axis denotes the nodes in HST-geo and the vertical axis denotes the normalized association probability. For example, "sky" has highly probability with the nodes covering the top area of an image and "horse" has highly probability with the nodes covering the middle area of an image. To qualitatively evaluate the association, for each noun attribute, we average the image patches assigned to it. Interestingly, as illustrated in Fig.4 (right), although learning in a weakly supervised way, our association shows the similar spatial priors of the object categories with [4] (see Fig.5 in [4]).

Fig.5 shows the image patches assigned to each noun are then split into multiple clusters according to the given adjectives. And we train a binary SVM classifier for each noun+adjective attribute based on those image patches using color histogram feature and SIFT bag-of-words feature.

5. Joint inference and learning

Take the learned HST-geo and association matrix Φ as an initialization, we infer $pt^+ = \{pt, A\}$ to simultaneously achieve the optimal scene configuration pt and attribute assignment $A = \{A^n, A^{adj}\}$, then re-estimate HST-geo and Φ. Thus rewrite Eq.2 as:

$$(\Theta, \Delta, \Phi)^* = \arg\max_{\Theta, \Delta} \log p(\mathbf{I}; \Theta, \Delta, \Phi) \qquad (8)$$

$$\propto \arg\max_{\Theta, \Delta} \sum_{m=1}^{M} \log \sum_{pt_m^+} p(I_m, pt_m^+; \Theta, \Delta, \Phi)$$

Table 1. The learning algorithm

Algorithm Iterative HST-att Learning
Initialization
1   Learn HST-geo (*optimize Eq.2*)
2   Pursue Φ and train appearance models based on HST-geo
Jointly learn HST-att (*optimize Eq.8*)
3   Jointly infer pt^+ with attribute localization (*optimize Eq.9*)
4   Update Θ and Δ in HST-geo (*optimize Eq.2*)
5   Update Φ and train appearance models based on pt^+ (*Eq.7*)
6   Repeat 3 - 5 until convergence

pt^+ is inferred from maximizing the joint probability $p(pt^+, I; \Theta, \Delta, \Phi) \propto \exp\{-E(pt^+, I; \Theta, \Delta, \Phi)\}$.

$$E(pt^+, I; \Theta, \Delta, \Phi) \qquad (9)$$
$$= \sum_{v \in V_{pt}^{OR}, v_i \in Ch(v)} E^{OR}(v_i|v) + \lambda_1 \sum_{v \in V_{pt}^T, a^n \in A^n} E^n(a^n|v)$$
$$+ \lambda_2 \sum_{a^n \in A^n, a^{adj} \in A^{adj}} E^a(a^{adj}|a^n) + \lambda_3 \sum_{v \in V_{pt}^T, a \in A} E^T(a|I_v)$$

where $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$ are the parameters balancing the energy terms ($\Lambda = \{0.7, 0.1, 2\}$ in this paper). The first term measuring the scene configuration prior is the same as Eq.4. The second term measures the noun attribute association:

$$E^n(a^n|v) = -\ln \Phi(a^n, v) \qquad (10)$$

The third term is designed to model the co-occurrence of a noun and an adjective attribute

$$E^a(a^{adj}|a^n) = -\ln p(a^{adj}|a^n) \qquad (11)$$

where $p(a^{adj}|a^n) = \frac{\sum_{m=1}^{M} \mathbb{1}[a^n \in A_m^n] \mathbb{1}[a^{adj} \in A_m^{adj}]}{\sum_{m=1}^{M} \mathbb{1}[a^n \in A_m^n]}$ encodes the compatibility between a noun and an adjective and can be counted from the given text phrases.

The forth term is an attribute specific data term which represented by the image features of the terminal node,

$$E^T(a|I_v) = -\frac{1}{|I_v|} \ln p(a|I_v) \qquad (12)$$

where $a = \{a^{adj}, a^n\}$ denotes the noun+adjective attribute, I_v is the image region occupied by v, $F(\cdot, \cdot)$ is a (strong) classifier learnt by SVM and $p(a|I_v)$ is given by $p(a|I_v) = -\frac{exp\{F(I_v, a)\}}{\sum_{a'} exp\{F(I_v, a')\}}$.

Because of the tree structure of HST and the linear form of Eq.9, the dynamic programming algorithm can be employed to infer the optimal parse tree with the attributes $(pt^+)^* = \arg\min_{pt^+} E(pt^+, I; \Theta, \Delta, \Phi)$.

Then based on the inferred pt^+, the HST-geo (*i.e.,* Θ and Δ) and HST-att (including the association matrix Φ and the appearance SVM classifiers) can be updated under the learning-by-parsing framework [17]. We summarize the entire learning procedure in Table.1, which contains two aspects. (i) Learn HST-geo and Φ based on the multi-scale segmentations as an initialization; and (ii) Re-estimate HST-geo and HST-att based on the joint inference.

cloudy sky
yellow field, mountain

green field, blue sky
white tower

birdseye buildings
overcast sky

white horse
green grassland

dust-haze sky, road
green trees
prespective buildings

blue sky, elephant,
yellow field

waterfall, black cliff
green cliff

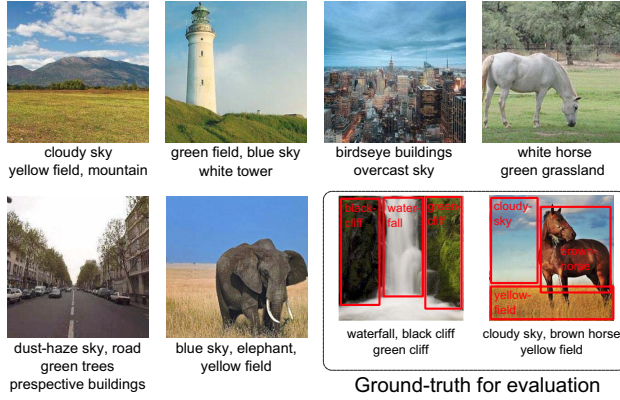cloudy sky, brown horse
yellow field

Ground-truth for evaluation

Figure 6. Examples of dataset and the ground-truth for evaluation.

6. Experiments

6.1. Datasets

There are two series of datasets relate to our task: scene datasets and image+text datasets. (i) For scenes, SUN dataset [18] contains 130,519 images with 397 categories. However, it is annotated at image level rather than specific regions. LabelMe Outdoor (LMO) [4] contains 2,688 fully annotated outdoor scene images. SUN Attribute database [7] contains 14,000 outdoor and indoor scene images with 700 attributes. In this paper, we focus on the outdoor scenes, while the indoor scenes are not treated because of the large 3D geometric variations caused by the view point changes. Moreover, some attribute types in [7], such as functions and affordances (*e.g.* playing, cooking), defined by human activities and recognized via human pose and activity reasoning, are beyond our scope. Therefore we select a subset of the above datasets for evaluation. (ii) For image+text, Kulkarni *et al*. [9] generate descriptions from scratch based on detected object, attribute, and prepositional relationships. Ordonez *et al*. [12] designed a SBU Captioned Photo Dataset through retrieving thousands of Flickr queries. Farhardi *et al*. [6] proposed the CORE dataset including 2,800 images with segmentations and attribute annotations for vehicles and animals. The above datasets provide a multitude of descriptions for images that are usually related to image content, however, they are not designed specifically for the natural scenes. Most of them focused on the objects, humans or the functional activities. Furthermore, those datasets always do not share intrinsic structures, in contrast, our goal is studying both the text descriptions and image configurations.

Therefore, we have created a new outdoor scene dataset as shown in Fig.6. The dataset (1226 images of 256×256 pixels in size) was selected from LMO [4] and SUN Attribute dataset [7]. To tolerate more objects (*e.g., wild animals*), we also added some images collected from Google images and Flickr, and got 12 categories in total. Text de-

scriptions were created by one author to ensure consistency and are publicly available.[1] Finally, we got the attribute set $\mathcal{A}^n$={*sky, flower, mountain, ibis, horse...*}, $\mathcal{A}^{adj}$={*blue, cloudy, rocky, snowy, brown...*} which contains 17 noun attributes and 30 noun+adjective attribute pairs in total. The average number of noun+adjective pairs attached to each image is 3. And for each noun+adjective pair, the average image number is 96. The dataset is split into 645 images for training (50 images per noun+adjective pair in average) and the rest for testing. For the testing set, we also ask people to localize the attributes through bounding box B_{gdth} as ground truth for evaluating the part localization accuracy, as it is shown at the bottom right panel of Fig.6.

6.2. Attribute Recognition

Baselines We first compare our method in attribute recognition, which evaluates the accuracy of an attribute presence in images. (i) *cKernel+SVM:* Xiao *et al*. [18] showed the combined feature kernels result in a significantly more powerful classifier than any individual kernel. We compare a combined kernel generated from gist, dense SIFT, HOG 2×2, self-similarity, and geometric context color histogram (see [18] for detail) and train a binary SVM classifier for each attribute. (ii) *BoW+SPM:* The spatial pyramid matching (SPM) proposed by Lazebnik *et al*. [16] partitions an image into increasingly finer spatial subregions and computes the SIFT bag-of-words (BoW) feature from each sub-region. (iii) *HST-geo:* To evaluate the contribution of attribute association, we also compare our method with HST-geo [17]. Specifically, for a given image, we first parse it from its multi-layer segmentation and classify each terminal nodes in the parse tree by the classifiers trained in (i).

Fig.7 shows the average precision (AP) for classifying each attribute and the mean average precision (MAP) for the entire attribute set is reported in Table.2. BoW+SPM shows lower performance because the lack of color feature which is a strong cue in scene attribute recognition. Though HST-geo and cKernel+SVM share classifiers, cKernel+SVM performs better because those classifiers are trained at the image level while the testing inputs of HST-geo are just image patches. Benefit from integrating scene geometry with attributes, our method outperforms all others.

6.3. Attribute Localization

Baselines For attribute localization, we benchmark our method against a fully supervised sliding window method (SW-FS) [10]. SW-FS trains an attribute classifier using ground truth bounding boxes as positive examples and random rectangles from each negative image for negative data. By treating localization as localized detection, the SW-FS applies attribute classifiers subsequently to sub-images at

---

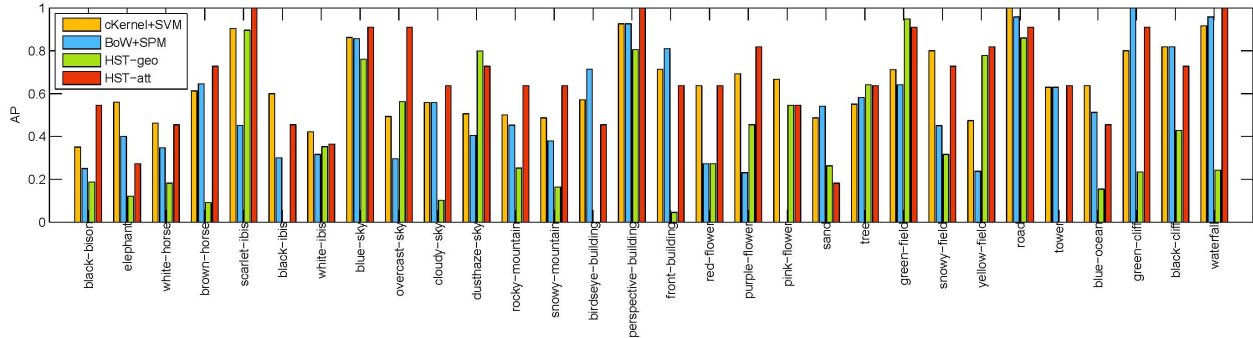[1]*http://www.stat.ucla.edu/~shuo.wang/SceneAtt.rar*

Figure 7. Average precision of attribute recognition.

different locations and scales. The detected sub-windows is ordered by the classification score and taken as indications for the presence of an attribute in this region by non-maximum suppression with 0.3 overlap threshold. In addition, we also compare with HST-geo for evaluating the attribute association.

Fig.8(b) shows the comparison of the benchmark methods with ours. Without the geometric constraint, (i) Certain attributes will be confused by appearance (*e.g.* HST-geo locates "sky" at the bottom region in the first row of Fig.8(b)), and (ii) The semantic region will be divided into fragments (*e.g.* the "black-bison" in SW-FS). Fig.8(a) shows the attributed parse trees and configurations generated from the joint inference and Fig.8(c) shows more localization results.

We quantitatively evaluate the attribute localization performance by following the procedure adopted in [18]. A ground truth bounding box (B_{gdth}) annotated "blue sky" implies if a localized bounding box (B_v) has at least $\mathcal{T}\%$ overlap with B_{gdth}, it can be correctly classified as "blue sky". Specifically, a correct localization has $\frac{area(B_v \cap B_{\mathrm{gdth}})}{area(B_v)} >= \mathcal{T}$. We do not care if the ground truth window is larger than the localization, *e.g.* a "blue sky" patch is correctly localized even if the ground truth "blue sky" has much greater spatial occupation. In this experiment, we set $\mathcal{T}\% = 50\%$. The threshold of 50% is set deliberately low to tolerate the inaccurate bounding box of highly non-convex objects, *e.g.* steep mountain. We use 11-point interpolated average precision [14] to evaluate the localization accuracy. The average precisions (AP) for each attribute are shown in Fig.9. The mean average precision (MAP) reported in Table.3 shows a surprising improvement of attribute localization of our method.

7. Discussion and future work

This paper presents a weakly supervised method for learning the scene configurations with attribute localizations. (i) We quantize the space of scene configurations by an Hierarchical Space Tiling (HST) and utilize a learning-by-parsing strategy to do parameter estimation; (ii) We discover the relationship between the scene parts and attributes

Table 2. The attribute recognition performance

	cKernel+SVM	BoW+SPM	HST-geo	HST-att
MAP(%)	64.48	53.11	51.67	**67.58**

Table 3. The attribute localization performance

	SW-FS	HST-geo	HST-att
MAP(%)	33.88	32.55	**50.22**

(nouns and adjectives) by an association matrix; (iii) We joint infer the scene configuration and attribute localization by dynamic programming. Our experiments show the promises in simultaneous parsing and localization. The attributes used in this paper are related to local object and regions, but there are also global attributes (style of the whole parse tree) such as aesthetics, which we are studying in ongoing work by extending our model to an attribute grammar.

8. Acknowledgement

References

[1] A.Oliva and A.Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001.

[2] B. Babenko, N. Varma, P. Dollar, and S. Belongie. Multiple instance learning with manifold bags. *ICML*, 2011.

[3] T. Berg and A. Berg. Automatic attribute discovery and characterization from noisy web images. *ECCV*, 2010.

[4] C.Liu, J.Yuen, and A.Torralba. Nonparametric scene parsing: label transfer via dense scene alignment. *CVPR*, 2009.

[5] D.Parikh and K.Grauman. Relative attributes. *ICCV*, 2011.

[6] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. *CVPR*, 2009.

[7] G.Patterson and J.Hays. Sun attribute database: discovering, annotating, and recognizing scene attributes. *CVPR*, 2012.

[8] J.Zhu, T.F.Wu, S.C.Zhu, X.K.Yang, and W.J.Zhang. Learning reconfigurable scene representation by tangram model. *WACV*, 2012.

[9] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: understanding and generating simple image descriptions. *CVPR*, 2011.

(a) Parsing outdoor scenes with attributes

(b) Attributes localization comparision
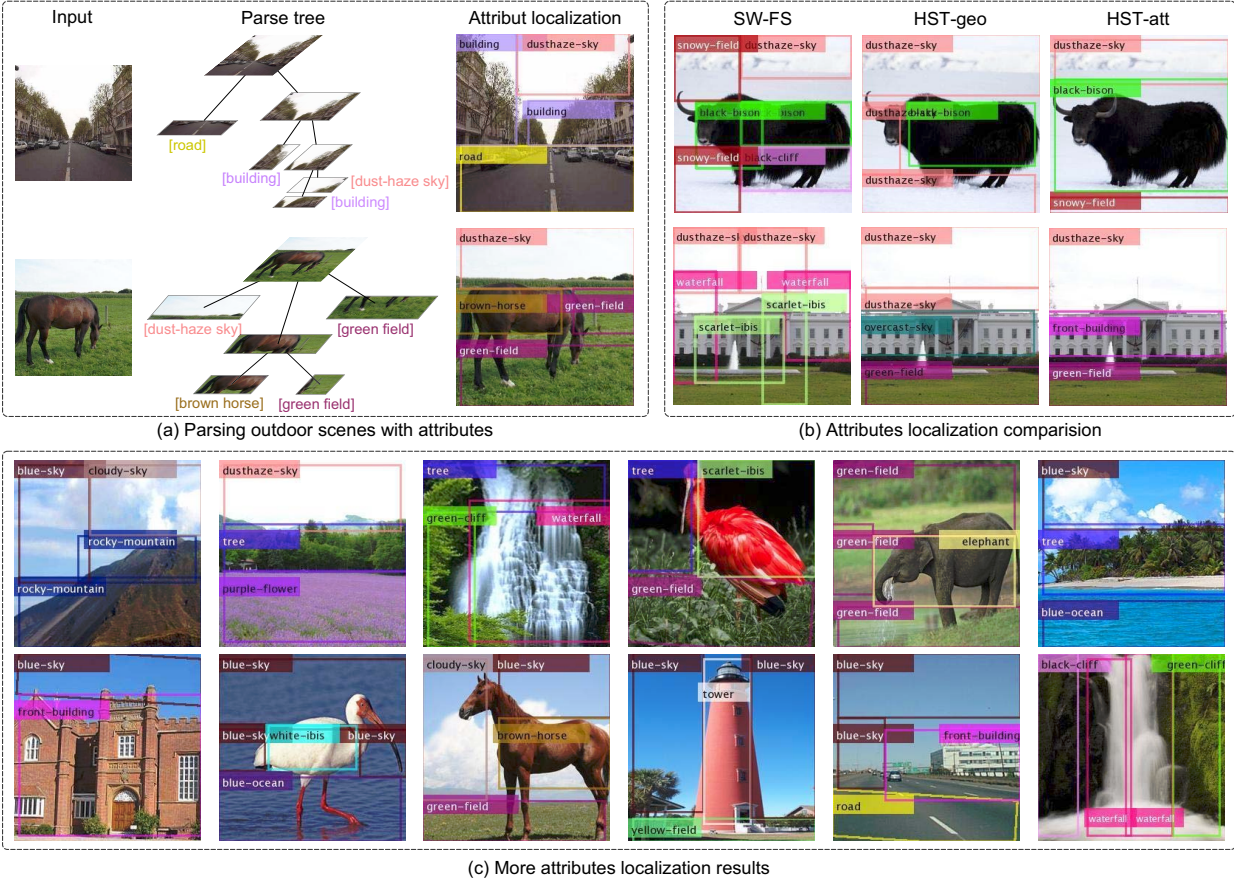
(c) More attributes localization results

Figure 8. Experiment results. (a) Parse trees with associated attributes. (b) Comparison of baseline methods and ours. (c) More attribute localization results from our method.
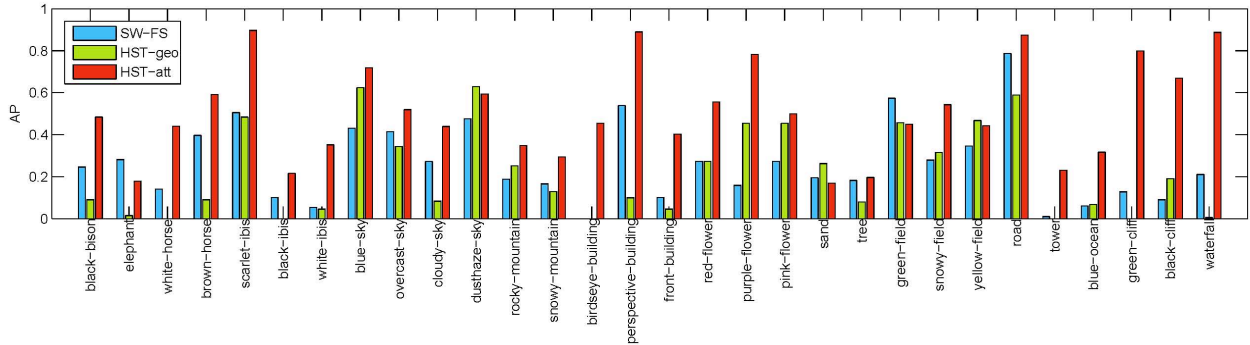

Figure 9. Average precision of attribute localization.

[10] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: object localization by efficient subwindow search. *CVPR*, 2008.

[11] L.Fei-Fei and P.Perona. A bayesian hierarchical model for learning natural scene categories. *CVPR*, 2005.

[12] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: describing images using 1 million captioned photographs. *NIPS*, 2011.

[13] P.Felzenszwalb and D.Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004.

[14] G. Salton and M. McGill. *Introduction to modern information retrieval*. McGraw-Hill, 1986.

[15] S.C.Zhu and D.Mumford. A stochastic grammar of images. *Found. Trends. Comput. Graph. Vis.*, 2006.

[16] S.Lazebnik, C.Schmid, and J.Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2006.

[17] S. Wang, Y. Wang, and S. C. Zhu. Hierarchical space tiling in scene modeling. *ACCV*, 2012.

[18] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *CVPR*, 2010.