Hierarchical Space Tiling (HST) for scene classification and attribute localization

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 representation named Hierarchical Space Tiling (HST) *and a* weakly supervised *method for simultaneously leaning the scene configurations and attributes from a collection of natural images associated with attributes in text, where the precise localization of each attributes left unknown. We start with an over-complete HST which quantizes the huge and continuous scene configuration space in an And-Or tree (AOT) through a small dictionary. Then estimate the HST model via a learning-by-parsing algorithm. Next, we iteratively pursue an association matrix for assigning attributes to the scene parts and re-estimate the HST model. Finally, given an image, we compute the most probable parse tree with the associated attributes as an instantiation of the HST by dynamic programming. We demonstrate the proposed by: (i) Representative and learning efficiency. We show the HST can approximate valid scene configurations with less errors using smaller number of primitives. (ii) Scene category classification. Considering only the geometry, we achieve the categorical scene configuration (priors) and further yield discriminative scene classification performance. (iii) Attribute classification and localization. By training the HST cross scene categories and exploring the relationship between the scene parts and attributes, we improve both the accuracy of attribute recognition and attributes localization.*

1. Introduction

If asked what the "prairie" looks like, one may describe it as a scene with "stretch of open, blue sky, green grassland,...". This description represents different visual cues, such as geometry and semantics, which are common to prairie and distinct from other scenes. A natural scene category usually shares geometric configurations and semantic attributes. The configurations explain the spatial layouts of a scene which are composed by the objects (*e.g.,* buildings, roads, vehicles) and regions (*e.g.,* sky, water, vegetation) of varying shapes. And the attributes, described by text, usually contains the nouns and adjectives, corresponding to semantic meanings of the objects/regions and their characteristics (*e.g.,* blue, green, smooth). Naturally, there are large variations in the representation of scenes depending on the image resolution. And the inside objects as well as their appearances are protean because of the changes in viewpoint, scale, illumination and season. A well-known representation that can explicitly address such complexity effectively is the family of hierarchical compositional models, which are reconfigurable and can generate combinatorial number of configurations through a small dictionary of shape elements. However, learning the structures of such models remains a challenge in vision, either for scene or object categories. Two main factors contribute to the difficulties of structure-learning: (i) The space of the internal nodes is huge or essentially continuous; (ii) The representations are often ambiguous (*i.e.,* not identifiable) and thus the learned model partially loses its power in parsing or classification as it diffuses probability over multiple possible interpretations.

In this paper, we propose a representation named *Hierarchical Space Tiling (HST)* and a *weakly supervised learning* method for simultaneously leaning the scene configurations and attributes from a collection of scene images associated with attributes in text, where the precise localizations of the attributes are left unknown. HST model contains two parts: HST-geo [2] and HST-att [1], modeling the geometric configurations and scene attributes respectively. We outline the three major issues in learning the HST model as following.

(i) HST-geo: Represents the scene configurations. We define a ground-truth *scene configuration* C as a label map with objects and regions, and define a *scene category* as an

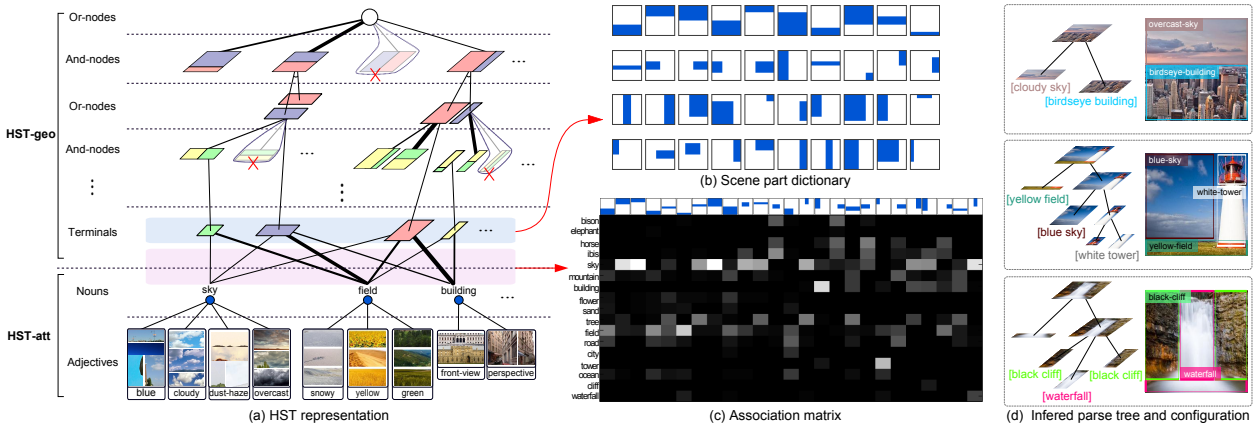

Figure 1. Illustration of HST. (a) The definition of HST model containing the HST-geo and HST-att. (b) Scene part dictionary which is formed by the terminal nodes (blue in (a)). (c) Association matrix which measures the assignable probabilities between scene parts and noun attributes (the red in (a)). (d) The inferred parse tree and the corresponding configurations with the attributes.

unknown set Ω^* of all valid configurations. Given training images, apply segmentation to achieve their configurations $\mathbf{C} = \{C_m : m = 1...M\} \subset \Omega^*$. The learning of HST-geo can equivalently be formulated as a stochastic grammar G. G is fully generative and its language is the set of all valid configurations: $\Omega(G) = \{C : C = g(pt; \Delta)\}$, where pt is the parse tree for C, Δ is the dictionary of the model, and $g()$ is the generation function. $\Omega(G)$ should cover all variations in $\mathbf{C}$ with high probability and thus be diverse, and also generalize well to the underlying set Ω^* to achieve reasonable parsing on the test dataset.

As shown in the top part of Fig.1(a), the HST-geo quantizes the huge space of scene configurations in an And-Or tree (AOT), where an And-node represents a way of geometric decomposition; an Or-node represents alternative decompositions with branching probabilities, and a terminal node is a shape element (*e.g.,* squares, rectangles) corresponding to the scene part. The terminal nodes of different sizes, locations and shapes form a hierarchical scene part dictionary (Fig.1(b)). The HST-geo can generate a combinatorial number of configurations (in the order of 10^{24}), which provides the potential to account for the variety of scene configurations.

(ii) HST-att: Represents the scene attributes. Beyond the geometric configurations, scene attributes are given by the text descriptions, consisting of nouns (*e.g.,* field, sky) and adjectives (*e.g.,* green, cloudy). They define the objects/regions inside a scene and their characteristics. The attributes can be represented as a two-level AOT (bottom part of Fig.1(a)), where each noun attribute is seen as an appearance-Or node having a mixture of adjectives, *e.g.,* the sky might be blue, cloudy, dust hazed or overcast. Each terminal node (scene part) in the HST-geo links to the noun attributes according to an association matrix (Fig.1(c)) which measures the co-occurrence of a local region and an object, *e.g.,* the grassland always appears at the bottom area of an image. Therefore, the scene configurations and attributes are integrate under a uniform framework.

(iii) Learning and inference. Owing to the design of HST model, we transfer the challenging *structure-learning* problem to a tractable *parameter-learning* problem. The learning includes two phases: (a) Learn HST-geo: starting with an over-complete HST-geo model, we update the parameters (branching probabilities at Or-nodes) while constructing the optimal parse trees (scene configuration) for each training images. (b) Learn HST-att: taking the learned HST-geo as an initialization, we iteratively obtain the attribute association matrix by non-maxima suppression and re-estimate the HST-geo. Finally, given an image, the optimal parse tree (scene configuration) can be inferred and the semantic attributes can be localized simultaneously by the dynamic programming (Fig.1(d)).

We evaluate the proposed method by showing: (i) Representative and learning efficiency. We compute the rate-distortion curves in coding theory, and show that our representation is clearly more effective than other popular representations. (ii) Scene category classification. Considering only the geometry, we apply the HST-geo to provide descriptions and show discriminative scene classification performance outperforms the state-of-the-art methods. (iii) Attribute classification and localization. By training the HST cross scene categories and exploring the relationship between the scene parts and attributes, we improve the accuracy of attribute recognition and the precision of localizing the attributes to image regions.

References

[1] S. Wang, J. Joo, Y. Wang, and S. C. Zhu. Weakly supervised learning for attribute localization in outdoor scenes. *CVPR*, 2013. 1

[2] S. Wang, Y. Wang, and S. C. Zhu. Hierarchical space tiling in scene modeling. *ACCV*, 2012. 1