Hierarchical Organization by And-Or Tree

Jungseock Joo*[a], Shuo Wang†[b], and Song-Chun Zhu‡[a,b]

[a]Department of Computer Science, UCLA
[b]Department of Statistics, UCLA

1 Introduction

A natural scene is composed of many components. See the example of the scene of beach in Figure 2. When we look at this image, our visual systems process a series of tasks in order to understand the whole scene. These tasks include to decompose the whole scene into parts, group them to form larger and larger parts, and organize discovered parts in a certain way. It has been a fundamental problem in computer vision to mimic these procedures by machine vision systems. However, this is a very challenging task due to the huge complexity arisen from an enormous number of distinct scene configurations, which are composed of a variety of objects and regions of varying shapes in different layouts.

In this chapter we will introduce a general model for scene or object categories that can represent varying configurations effectively. The desired proper-

*joo@cs.ucla.edu
†shuoshuow@ucla.edu
‡sczhu@stat.ucla.edu

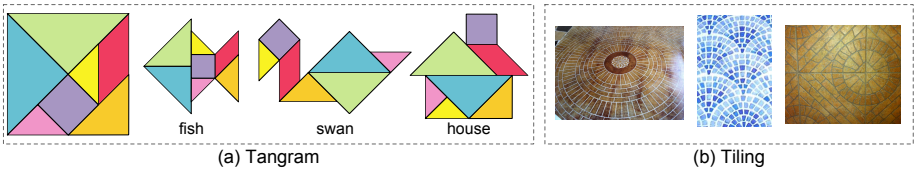


Figure 1: (a) The "Tangram", the ancient Chinese puzzle which consists of seven pieces, and a few examples of completed shapes formed by the pieces. One can composite an enormous number of different shapes by assembling the same set of pieces. (b) A various types of tilings, also called *tessellation*, in the real world. Although building blocks are simple and may be even identical, high order patterns can still emerge from specific configurations, namely, organizations.

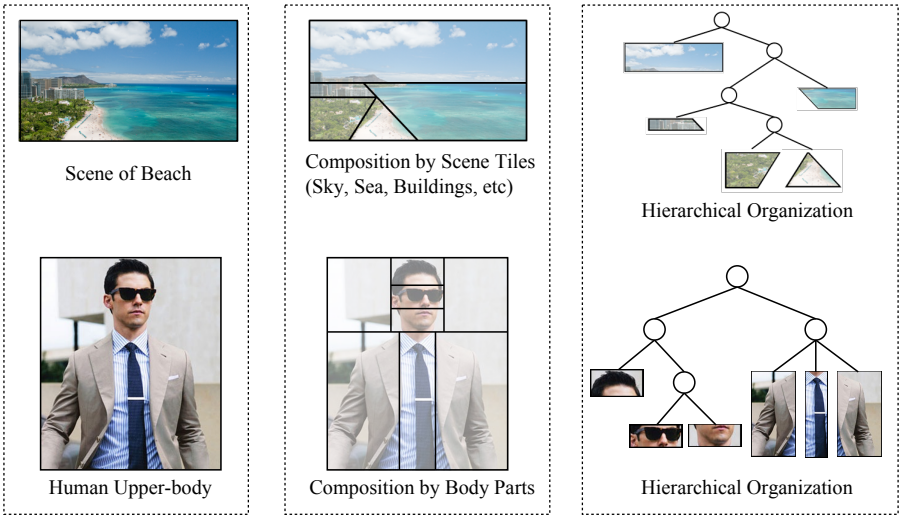


Figure 2: A natural scene (*top*) as well as an object (*bottom*) contain a number of components and their sub-components. We can completely understand the image by decomposing the whole into the parts and organizing them.

ties of such models can be summarized as follows:

i. It should incorporate **generic** grouping rules among image primitives at low-middle level interpretation (*i.e.*, Gestalt Laws) as well as **category-specific** production rules of parts at high level (*i.e.*, image grammar).

ii. **Compositionality** is required as it ensures that the model can be expressive enough to deal with hugely varying configurations of many components by a relatively small dictionary.

iii. The structural representation should be **flexible** so that it can adaptively capture unique configuration of each instance at multi-scales, as opposed to fixed representations.

iv. Finally, the learned models should be **unambiguous** and allow only one interpretation to each instance of a given scene or an object.

In order to fulfill such requirements, the proposed model will be a hierarchical compositional model based on the tiling method. The tiling, as shown in Fig. 1, can be seen as a process of composing complex shapes by assembling smaller and simpler parts. Fig. 1(a) shows a tiling puzzle, an ancient invention in China, called "Tangram". While it is composed of a small set of very simple pieces, one can composite an enormous number of a variety of complex shapes by assembling them. The same intuition can be also found in real world examples

such as tessellated street pavement and ceramic tile flooring. In such cases, one can observe complex high order patterns emerging from one or few types of tiles according to specific configurations, namely, organizations of tiles.

Inspired by these examples, each individual component of scene or object will be treated as a **tile** in the proposed model whose visual dictionary will be a collection of all observable tiles. Each tile is treated as a template that explains a specific part of the image. Then, the task of understanding the whole scene will simply become **tiling**, which is identifying proper tiles and assembling them. As the nature of tiling, we consider the assembly of tiles in 2D space in this chapter, in contrast to another class of models to cope with 3D arrangement of parts or primitives.

Our framework which utilizes image parts (tiles) and their relations is closely related to a series of theories in part-based object recognition of human vision, e.g., "Recognition-by-Components" by Biederman [1]. According to these models, humans perceive given scenes as their "structural descriptions" with a limited set of known components in memory while a huge flexibility is achieved through combinations of the components. On the other hand, another class of theories, "image-based" models [5, 21], suggest that our brains store many viewpoint-specific images of the same object. By analogy, our model also incorporates multiple templates, each of which explains an aspect specific to viewpoint or appearance type. Such treatment allows us to deal with complex and nonrigid parts of real-world objects such as humans. In contrast to image-based models, we define the set of templates at the part level (rather than at the entire image level) and parse the image into the parts with selected templates where the relations among the parts are also captured by the model structure. Therefore, our proposed model can be seen as a combined approach that can benefit from both classes of models.

2 Background Review

In this section, a group of related researches on perceptual organization will be briefly reviewed. In particular, we will consider two different dimensions: 1) whether their grouping rules and parts are generic or category-specific (Sec. 2.1), and 2) whether their representations are built on a flat layer or in hierarchy (Sec. 2.2).

2.1 Grouping Rules: Generic vs. Category-Specific

At low level, an image can be seen as a collection of simple image features or primitives such as line segments, junctions, and so on. At this level of abstract, relationship among primitives is disregarded. It is the role of perceptual organization that exploits such relationship and detects the groupings of elementary primitives. Gestalt laws such as proximity, continuity, etc explain certain patterns of grouping capabilities of humans, which lead to advanced interpretation enriched by geometric contexts among primitives as middle level representation.

These grouping rules and simple primitives are *generic* and commonly observed in any types of objects and scene categories. The generic grouping rules of image structures have been studied in many works in the literature, including Lowe's early work [13]. Lowe viewed that the goal of perceptual organization is to find out image relations arisen from actual structure in the scene. He measured this quantity for each grouping rule, such as collinearity and parallelism. Mohan and Nevatia [14] also exploited such grouping rules to detect geometrically related edges for scene segmentation. These generic grouping rules often form simple and common groupings of primitives, such as L-junction. More recently, Wu *et al.* [24] defined a set of common "graphlets" (simple primitives and junctions) as basic building blocks and parsed the whole scene from detected graphlets in a bottom-up manner.

Beside generic parts, any object or scene class also has its own unique parts as well as distinct grouping rules, which can be seen as *category-specific* information. Thus, it is difficult to understand the entire pattern of image solely by generic rules. Such unique parts, which could be formed from generic parts, may have complex structures (compared to simple primitives) and be shared by objects within one or a few classes. Therefore, learning and representing them can't be achieved in the same way as generic parts and grouping rules. In late 1970's, Saund was among the first to go beyond generic Gestalt's laws [19]. He pointed out that the domain specific knowledge plays an important role in shape representation and one might lose this important information when relying on a fixed set of generic shape primitives alone.

More generally, the goal of many high level vision tasks is to learn category-specific dictionaries of parts and their configurations. These dictionaries tend to contain more complex elements than common primitives so that they can reflect distinct properties of each category of object or scene. The corresponding configurations can also capture unique structure or relations of parts. For example, a human and a dog have different sets of parts and different configurations, and none of them can be identified by generic rules without domain knowledge.

2.2 Organization: Flat vs. Hierarchical

The generic grouping rules, such as Gestalt laws, have been often posed as relational constraints on the parts, which are modeled in a flat layer. For example, Zhu [28] proposed a mathematical framework based on Markov Random Field (MRF) whose neighborhood structures captured relationship between line segments. Through the structures, Gestalt laws were explicitly modeled as pairwise features so that they could act as constraints posed on shape elements. Porway *et al.* [16] also employed MRFs for aerial image parsing where the common elements of aerial images such as parking lots, roads, etc were defined on the graph. Then, the statistical constraints such as relative position were added between objects.

However, certain relations or groupings can be better organized and expressed in the hierarchy of different levels of abstract. A fractal pattern is a good example in which one can observe the law of symmetry recursively at in-

finite scales. Let's also recall the beach example in Fig. 2 which contains many components and their subcomponents. One can easily imagine a huge complexity that would be caused by modeling all components and their relations together on the flat representation.

The use of hierarchical representation for image modeling dates back to 1970s in Fu's early works [9], namely syntactic approaches where pattern structures and sub-pattern relations were modeled as symbolic tokens and production rules by analogy to natural languages. Dickinson *et al.* [4] adopted a hierarchical Bayesian network for 3D object recognition, where layers of short boundaries, object faces, and aspects were linked hierarchically. Sarkar and Boyer [18] also used the Bayesian network for grouping primitives into hierarchical structures in aerial images. In both models, groupings were governed by conditional probabilities defined over layers in the hierarchy. More recently, Geman and collaborators [2] presented grammatical and compositional frameworks with applications such as vehicle license plate recognition [10]. Zhu and Mumford [29] also proposed a general framework for image grammar named And-Or Graph, which we adopt in our model and will discuss in details in Sec. 3.

The key advantage of these approaches is that they can represent an enormous number of distinct configurations by composing a relatively smaller number of elements, instead of enumerating all possible configurations. Also, hierarchical structures further allow us to limit local complexity at each scale. As discussed at the beginning of this chapter, these are critical aspects in modeling highly complexed and versatile scene or object classes.

Again, the remaining question is how to learn image parts and their relations. In the rest of this chapter, we will introduce a hierarchical compositional model based on "Hierarchical Tiling". In this model, the grouping rules will be defined by region-based recursive decomposition and each sub-region will correspond to an atomic element in the dictionary (Sec. 3). Then the learning problem can be posed as node pruning and parameter estimation problem (Sec. 4).

3 Hierarchical Organization by AOT

Now, we provide the definition and details of our model for hierarchical organization. We adopt And-Or Tree (AOT) [29] as our main framework. AOT has been used for modeling objects and scenes in the literature of computer vision [27]. An AOT, as the stochastic image grammar, represents the hierarchical decompositions of elements and produces a number of varying configurations by alternating sub-components subject to probabilistic distributions defined over nodes and edges.

Each node in AOT plays a distinct role according to its node type. As Fig. 3. illustrates, an AOT has three types of nodes: AND nodes, OR nodes and Terminal nodes. Note that all nodes are associated with specific sub-regions and the root node corresponds to the whole region of image. Each type can be characterized as follows.

i. An **AND** node represents the composition of two sub-regions. For instance,

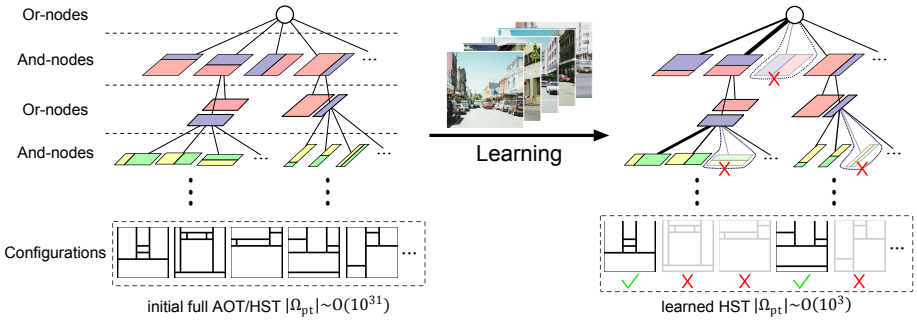


Figure 3: During the learning process, a number of invalid configurations are eliminated from the initial model. This results in a huge drop in the complexity of the model and the final model only contains a compact set of meaningful configurations which can be frequently observed in the training images.

'upper-body' = 'head' ∪ 'torso'. By the definition of hierarchical tiling, AND nodes always have two child nodes.

ii. An **OR** node contains several alternating ways to decompose the current region. This is a switch of indicating how and where to partition the current region.

iii. A **Terminal** node corresponds to the most elementary region which is not decomposed further.

Note that an AOT is a 'whole' representation of the entire scene class, in the sense that all possible decompositions of all sub-regions are integrated in this AOT. In order to represent a particular image, one needs to make choices at OR nodes to select specific decompositions. We call this process *parsing*, which yields corresponding representations as follows:

i. A **Parse Tree** is an image-specific instance drawn from AOT. This is a set of selected nodes including terminal and non-terminal nodes.

ii. A **Configuration** means a spatial layout of elementary regions in a parse tree. In other words, it is a set of terminal nodes in a parse tree, which does not reflect hierarchical relationship.

iii. Then, a whole **AOT** can be seen as an entire collection of all possible parse trees and configurations.

One important benefit of this representation is flexibility which is required to account for varying scene components and configurations. When built on 8 by 8 grid, the AOT can generate more than 4×10^{31} different parse trees. This flexibility comes from only 1,296 rectangular building blocks which are re-configurable. The efficiency of this model partly relies on the fact that smaller

sub-regions - nodes in lower layers in AOT - can be shared by multiple parent regions at higher order. However, such a huge flexibility also introduces a counter effect on increased complexity and ambiguity. We will discuss this issue in the following section in details.

3.1 Mathematical Formalism

In this subsection, we define notations and introduce mathematical formalisms.

Given a set of N training images $\{I_i\}$, our objective is to learn an AOT with visual dictionary and associated parameters. Let us define the AOT as follows.

$$AOT = (S, V; \Theta, \Delta) \tag{1}$$

where S is a start symbol at root, *i.e.*, the whole region, V is a set of nodes in AOT. A node, $v_i \in V$, has one of types: {AND, OR, Terminal} as section 3. Θ is the set of model parameters which control the frequencies of decomposition rules being activated at OR nodes. The tiling dictionary of the scenes is denoted by Δ, which is also a set of terminal nodes in V.

From the AOT, the learning problem can be formulated as maximum likelihood estimation (MLE).

$$(\Delta, \Theta)^* = \arg\max_{\Delta,\Theta} \sum_{i=1}^{N} \log p(I_i; \Delta, \Theta). \tag{2}$$

In the AOT model, each image I is generated by a hidden parse tree, pt. Then the data likelihood in Eq. (2) can be marginalized over parse trees and further factorized as follows.

$$p(I_i; \Delta, \Theta) = \sum_{pt} p(I_i, pt; \Delta, \Theta) \tag{3}$$

$$= \sum_{pt} p(I_i|pt; \Delta) \cdot p(pt; \Theta). \tag{4}$$

For a certain parse tree, pt, the first factor of the product in eq. (4) represents the likelihood of an image given the parse tree. In order words, it measures how well the parse tree and corresponding configuration is suited to or explains the given image. And the second part, $p(pt; \Theta)$, is a *prior* probability of the parse tree and this measures how commonly this parse tree would be used. This part is not affected by the choice of image.

4 Learning AOT

So far, we have discussed the general structure of our model. The next step is to learn actual models from training images. To learn a model means to define the whole structure and estimate optimal parameters such as probabilistic distribution, from training data. In our model, this can be understood as learning

how frequently each decomposition has happened and ruling out those paths that has never or rarely occurred.

This procedure can be easily understood when we think of what we do for learning our visual world. For example, let's imagine a typical scene of 'beach' (as one presented in Fig. 2). Then one would probably draw in mind a horizontally divided scene with the sky at top and ocean at bottom, because this spatial configuration is very common in beach scenes that we have observed, and we have learned and stored such frequency of configurations in our mind.

Therefore, our learning procedure follows the exact same strategy as humans. The algorithm takes as input a set of training images and infers the most probable interpretations of them, *i.e.*, parse trees and configurations. Then, it can evaluate what kinds of configurations are the most common and how frequent each one is. Such information are stored as parameters of the learned model, and eventually, can be used for analyzing a new image.

On the other hand, the main difficulty in many structure learning algorithms comes from the fact that there are too many different ways in decomposing the scene into parts, *i.e.*, ambiguity. This difficulty can be alleviated here by constraining the feasible set of structures by definition of hierarchical tiling described in the previous section. The hierarchical tiling AOT contains a number of rectangles on the grid as basic building blocks as well as rules of decomposition. In this representation, the original continuous geometric space is quantized at the resolution of the grid, and moreover, factorized into the local forms of three regions: one parent region at an AND node and two subsequent subregions. Therefore, the complexity is locally limited, and this makes the model manageable in learning. Note that, despite this constraint, it can still represent a combinatorial number of parse trees, which provide enough flexibility to modeling a variety of configurations.

Figure 3 illustrates the key idea of the learning procedure which can be seen as a shrinking process. It first establishes a very "fat" and highly over-complete initial model. This model can generate an exponential number of different configurations. Some of these configurations are useful (they correspond to the real examples of natural scenes), however, most of the other configurations do not make any sense so they are unable to capture meaningful structure of any natural scenes. Therefore, those meaningless configurations will be gradually eliminated from the initial model during the learning procedure. Then eventually, the learned model can generate a much more compact set of configurations and parse trees, which one can commonly observe in real images.

4.1 Iterative Learning

In our formulation, a parse tree is a latent variable which is not observable. One common algorithm used for maximum likelihood estimates with latent variables is Expectation-Maximization (EM) algorithm [3]. This is an iterative algorithm and alternates between evaluating the posterior distribution of latent variable and updating model parameters, based on the current estimates at each iteration. Our learning algorithm follows a similar iterative strategy which alternates

between inference of the optimal parse trees and updating parameters. The details of each step can be summarized as follows.

1. **Inference**. Inference is the task of evaluating the most probable parse tree which can be considered as the best interpretation of a given image under the current parameters of AOT. We obtain the optimal parse tree for each image by dynamic programming (DP) in a bottom-up process. For a given image I_i, the optimal parse tree, pt_i^*, maximizes following probability.

$$pt_i^* = \arg\max_{pt} p(I_i|pt;\Delta^t)\cdot p(pt;\Theta^t). \tag{5}$$

   The parse tree prior is the product of branching frequencies at OR nodes.

$$p(pt;\Theta^t) = \prod_{v\in V^{OR}\subset pt} \Theta^t_{(v,v_{ch})}, \tag{6}$$

   where $\Theta^t_{(v,v_{ch})}$ is the branching frequency from an OR node v to its child node, v_{ch}.

2. **Activation Frequency Update.** After obtaining the optimal parse trees, now the parameters of model are updated. These parameters include the activation frequency, Θ, which indicates frequencies of decomposition rules.

$$\Theta^{t+1}_{(v,v_{ch})} = \frac{\sum_i \mathbf{1}[v, v_{ch}\in pt_i^*]}{\sum_i \mathbf{1}[v\in pt_i^*]}. \tag{7}$$

3. **Node Pruning.** According to the updated frequency, the dictionary is compressed by pruning nodes which have been never or rarely activated.

$$\Delta^{t+1} = \Delta^t\backslash\{v; f(v)<\epsilon, v\in\Delta^t\}, \tag{8}$$

$$f(v) = \frac{1}{M}\sum_i \mathbf{1}[v\in pt_i^*]. \tag{9}$$

These steps are repeated until the model converges. At the beginning, an initial AOT contains a huge number of decomposition rules and a large size of the dictionary and there exists a very high ambiguity on parsing images. As iterations proceed, the model parameters keep being refined, also the size of the dictionary becomes smaller. A series of relevant experimental results will be presented in following sections with applications to the scene and the human body.

5 Case Study I. Scene

In this section, we present a concrete development of the introduced algorithm for scene modeling and its evaluation. The experimental results of this section was reported in [23].

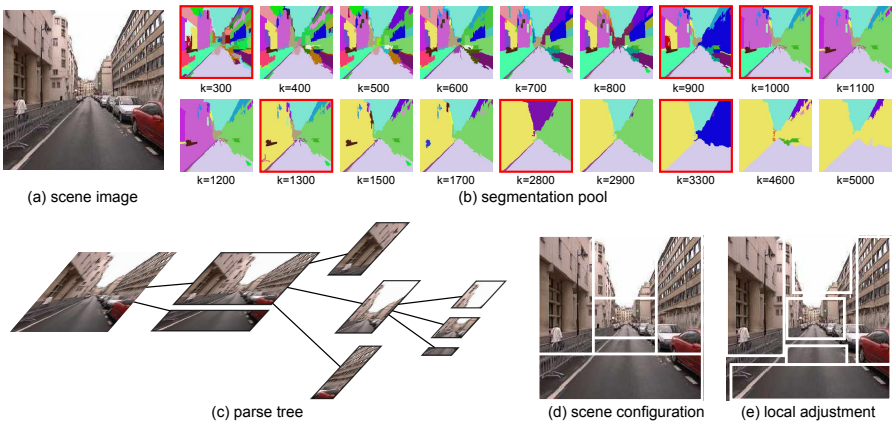


Figure 4: Parse an image into scene configuration. (a) Input image. (b) Segmentations in different layers. (c) The optimal parse tree of given image. (d) Scene configuration. (e) Scene configuration with localized parts.

For the purpose of scene analysis, a dataset of natural scene images has been proposed to computer vision community [17]. This dataset contains $2,688$ images from 8 categories of outdoor scenes including coast, highway, open-country, street, forest, tall-building, inside-city and mountain. Figure 5 shows the examples of each category in the dataset.

For each image, our algorithm first generates multiple segmentations by graph-based segmentation method [7], as shown in Figure 4 (b), while varying the parameter, k, which controls the granularity of segmented regions. From the set of segmentation layers, we obtain the optimal parse tree and corresponding configuration, which are consistent with learned parsing prior and preserve the homogeneity of each terminal tile. That is, we encourage the model to parse an image into a more familiar configuration where each perceptually homogeneous sub-region, an image segment, is explained by an elementary part in one piece.

5.1 Qualitative Results

Table 5.1 shows the statistics on the complexity of AOT. The size of parsing space that an initial AOT defines is combinatorial. It contains a huge number of region decomposition rules and this can generate an enormous number of distinct parse trees. This also implies a high ambiguity. Through the iterative learning procedure, the admissible parsing space quickly shrinks by pruning many infrequent parsing rules and nodes. After convergence, the learned AOT only contains a compact set of common parsing paths and nodes.

5.2 Scene Category Classification

The goal of scene category classification is to predict a scene category to which each image belongs. This is a multi-class classification problem which has attracted many researches in computer vision. Many prior works have focused

Table 1: The shrinkage of AOT for "street" scene at each iteration round

Round	$\|V^{AND}\|$	$\|V^{OR}\|$	$\|V^{T}\|$	$\|\Omega_{pt}\|$
0	6048	1296	1296	4.48×10^{31}
1	570	519	366	8.01×10^{7}
2	351	386	256	2.23×10^{5}
3	238	302	184	1.14×10^{3}
4	221	290	173	9140

on exploring better image feature without considering structural representation (Gist [15], Bag of Words [12]), or building their models on limited or fixed structure (spatial pyramid [11]). In contrast, our model can take advantage of much more flexible representation by AOT.

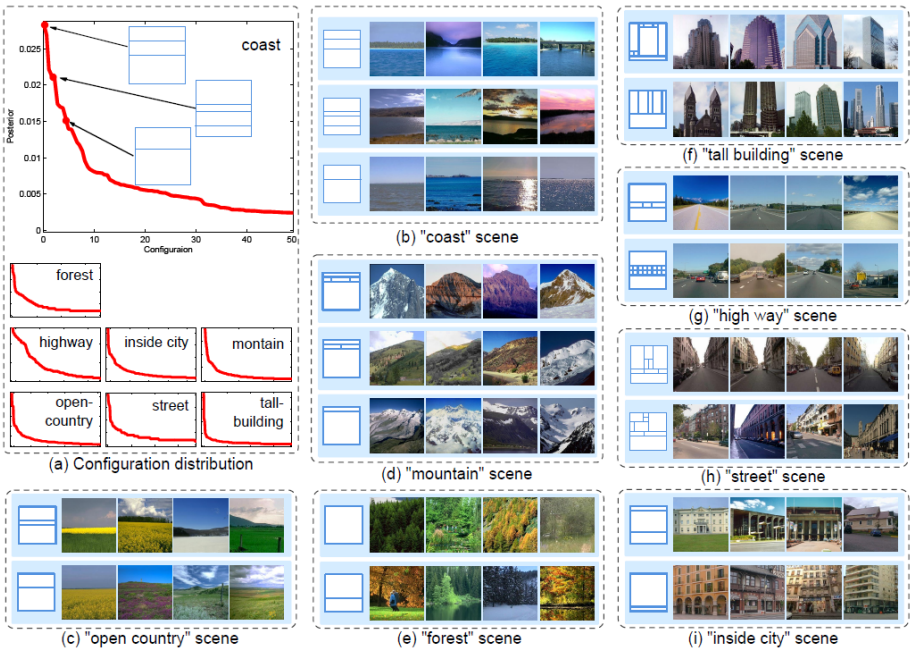
Specifically, we obtain a set of typical configurations of each scene category from the learned AOT, as shown in Figure 5. We use SIFT descriptors and color moments of each terminal window as image features and train category classifiers by support vector machine (SVM). Given a test image, we assign the best category whose prediction score is maximum. Note that the ground-truth segmentations (label map) of training images are provided in this dataset and we used them for compatible comparisons with the other method.

We compare the performance of our model with prior works including: (i) a holistic "Gist" feature based method [15], (ii) a BoW based method [12], (iii) spatial pyramid matching (SPM) method [11], and (iv) an extension of SPM named locality-constrained linear coding (LLC) [22]. (iv) the tangram model (Tgm) [25]. Figure 5 shows the average precision (AP) of different methods, where our method outperforms the others.

This can be a strong evidence supporting the needs of flexible and hierarchical models in understanding the scene. Without such a hierarchy, one can still identify some common visual words (BoW), but it loses the spatial information and the relationship among parts and fails to capture the context on the entire scene. Although some uniformly predefined configurations have been used in SPM, it still results in a poor performance. One possible explanation is that their configurations, regular grids at multi-resolution, are not coherent with real images of scenes. Therefore, by pursuing meaningful spatial layout from training data, the hierarchical tiling model can improve the classification performance.

6 Case Study II. Object: Human Figures

In this section, we present the application of our algorithm to an object with examples of human bodies. As shown in Figure 7, a complete human body can be understood as a hierarchical organization of body parts. In fact, such type of hierarchical models have been used for tasks such as human pose estimation [26] and general object detection [6] in the recent literature. The common idea



Methods	Gist[15]	BoW [12]	SPM [11]	LLC [22]	Tgm [25]	Ours[23]
AP(%)	72.15	84.57	84.92	87.97	86.07	**91.71**

Figure 5: Scene classification based on the categorical typical configurations. (a) The learned configuration distributions where the horizontal axis is the index of configuration and vertical axis is the posterior probability. (b)-(i) The categorical typical configurations for each scene category. The performance of scene classification is shown in the bottom table.

Figure 6: Examples of upper-bodies of human.

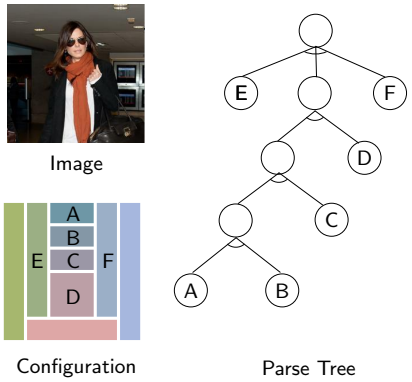


Figure 7: Similar to the case of scene, one can interpret the human body as a collection of body parts which can be organized in the object-level And-Or Tree. The configuration will be governed by the pose of body as well as different clothings or accessories (*jean, skirt, etc*) that each person wears.

behind such methods is to decompose the whole object into its parts and analyze them. Compared to the conventional whole template based approach without part definition, the part based approaches have their strength in capturing individual geometric variation of each part and relationship among parts, which has leaded to the improvement in performance [6].

While a majority of related works focus on learning parameters of manually defined structures of objects, there has been the other line of researches pursuing learning the unknown structure of objects from image [27, 8]. Our learning method introduced in this chapter also falls into this category. Then the task of learning is equivalent to identifying the hierarchical dictionary of body parts including varying types of appearance from raw images.

Figure 6 shows examples of input training images which contain upper-bodies of human. These images are collected from Internet for the experiment. Images are preprocessed by cropping and aligning with respect to the positions of head and waist.

The algorithm starts by learning appearance models at each rectangular sub-region in AOT. This is essentially a task to learn the conditional image likelihood given a terminal node, $p(I|pt)$ in Eq. (4). To model the likelihood of appearance, Hybrid Image Template (HIT) has been used in this experiment. HIT is a generative image model having four different types of low-level features: {sketch, color, texture, flatness}. Details can be found in [20]. A single HIT

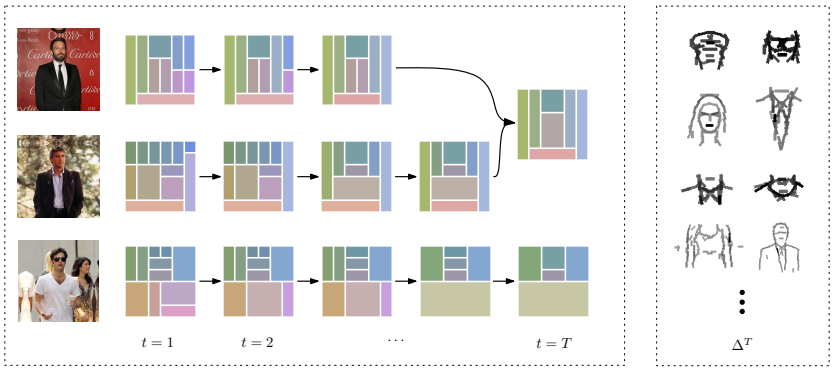


Figure 8: (*left*) The optimal configurations pooled from AOT at each iteration. At the beginning, the ambiguity is very high. As the learning proceeds, the optimal configuration becomes more meaningful, and finally captures the correct parts of human bodies. (*right*) Some popular elements in the dictionary of AOT after learning.

template can be learned for each terminal node to represent an individual part. Then, the whole AOT which contains a number of HIT templates can generate many compositional human poses, each of which is a composed HIT template for human body.

For each sub-region at a terminal node, the corresponding patches of all training images are cropped and clustered by their appearance into k distinct groups. Then, from each cluster, a single HIT template is learned so that one rectangular sub-region has k different appearance models. These k templates address the different appearance types of a part. For example, a head can be modeled as a mixture of templates including 'head with hat', 'head with long hair', and so on. For this reason, now the task of parsing includes the choice of specific appearance type at previous terminal nodes. Therefore, such a terminal node becomes another OR node whose children are a set of appearance templates. A complete parse tree now includes spatial configuration as well as associated appearance types of parts.

6.1 Learning Body Parts

At this point, we still do not have a clue of what sub-regions are true human parts and all templates are treated as potential parts. As the case of scene, we build a fat initial AOT and again go through iterations in order to develop and refine a compact model where ambiguous parts are suppressed.

Figure 8 shows a series of optimal configurations being developed through iterations. At the beginning, the ambiguity is very high as there are too many redundant parts and a prior on parse tree is yet immature. As the learning proceeds, the optimal configuration becomes more meaningful, and finally captures the correct parts of human bodies. Some of those parts are presented in

Figure 8. These are the most frequently used parts during the learning, and thus included in the learned dictionary.

From the result, one might wonder why the model can learn true parts or which parts are preferred over the others. There are two factors in deciding the optimal parse tree: the image likelihood from selected appearance templates and the parse tree prior which controls the overall frequency of parts being activated. The set of true atomic parts which can be modeled by rigid templates tend to be more robust from articulation, which leads to higher image likelihood. Therefore, we can think that some good appearance templates (hence, good parse trees) are more likely to be selected at earlier stages of learning and the other ambiguous parse trees will move toward a smaller set of good parse trees to which stronger priors are given.

7 Conclusion

In this chapter, a hierarchical representation for images and its learning algorithm were discussed. And-Or Tree (AOT) was adopted as the main framework to model the hierarchy of image structure. An algorithm to learn the parameters and dictionary of AOT has been suggested with mathematical formalisms. Finally, to demonstrate the introduced model and learning method, two concrete cases for natural scenes and human bodies have been presented with a various experimental results.

Acknowledgements. This work was supported by NSF CNS 1028381, DARPA MSEE grant FA 8650-11-1-7149 and MURI grant from ONR N00014-10-1-0933. We would like to thank Johan Wagemans and two anonymous reviewers for their valuable comments.

References

[1] Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[2] Elie Bienenstock, Stuart Geman, and Daniel Potter. Compositionality, mdl priors, and object recognition. In *Advances in Neural Information Processing Systems*, pages 838–844, 1996.

[3] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[4] Sven J. Dickinson, Alex P. Pentland, and Azriel Rosenfeld. From volumes to views: An approach to 3-D object recognition. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 55(2):130–154, 1992.

[5] Shimon Edelman and Heinrich H. Bülthoff. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, 32:2385–2400, 1992.

[6] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell*, 32(9):1627–1645, 2010.

[7] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, pages 167–181, 2004.

[8] Sanja Fidler and Ales Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[9] King-Sun Fu. *Syntactic Methods in Pattern Recognition*. 1974.

[10] Ya Jin and Stuart Geman. Context and hierarchy in a probabilistic image model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2145–2152, 2006.

[11] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[12] Fei-Fei Li and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[13] David G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, MA, USA, 1985.

[14] Rakesh Mohan and Ramakant Nevatia. Perceptual organization for scene segmentation and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(6):616–635, 1992.

[15] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[16] Jake Porway, Qiongchen Wang, and Song-Chun Zhu. A hierarchical and contextual model for aerial image parsing. *International Journal of Computer Vision*, 88(2):254–283, 2010.

[17] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.

[18] Sudeep Sarkar and Kim. L. Boyer. *Computing Perceptual Organization in Computer Vision*. World Scientific, 1994.

[19] Eric Saund. Putting knowledge into a visual shape representation. *Artif. Intell.*, 54(1):71–119, 1992.

[20] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (HIT) by information projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1354–1367, 2012.

[21] Michael J. Tarr and Heinrich H. Bülthoff. Image-based object recognition in man, monkey and machine. *Cognition*, 67:1–20, 1998.

[22] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[23] S. Wang, Y. Wang, and Song-Chun Zhu. Hierarchical space tiling in scene modeling. In *Asia Conference on Computer Vision*, 2012.

[24] Tianfu Wu, Gui-Song Xia, and Song-Chun Zhu. Compositional boosting for computing hierarchical image structures. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[25] Jun Zhu, Tianfu Wu, Song-Chun Zhu, Xiaokang Yang, and Wenjun Zhang. Learning reconfigurable scene representation by tangram model. In *WACV*, pages 449–456. IEEE, 2012.

[26] Long Zhu, Yuanhao Chen, Chenxi Lin, and Alan L. Yuille. Max margin learning of hierarchical configural deformable templates (HCDTs) for efficient object parsing and pose estimation. *International Journal of Computer Vision*, 93(1):1–21, 2011.

[27] Long Zhu, Yuanhao Chen, and Alan L. Yuille. Unsupervised learning of probabilistic grammar-markov models for object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):114–128, 2009.

[28] Song-Chun Zhu. Embedding gestalt laws in markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1170–1187, 1999.

[29] Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2(4):259–362, 2006.