A Stochastic Grammar of Images

Full text available at: http://dx.doi.org/10.1561/0600000018

A Stochastic Grammar of Images

Song-Chun Zhu

University of California
Los Angeles
USA

sczhu@stat.ucla.edu

David Mumford

Brown University
USA

David_Mumford@brown.edu

the essence of knowledge

Boston – Delft

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is S.-C. Zhu and D. Mumford, A Stochastic Grammar of Images, Foundations and Trends® in Computer Graphics and Vision, vol 2, no 4, pp 259–362, 2006

ISBN: 978-1-60198-060-1

Foundations and Trends® in Computer Graphics and Vision
Volume 2 Issue 4, 2006

Editorial Board

Editorial Scope

Foundations and Trends® in Computer Graphics and Vision will publish survey and tutorial articles in the following topics:

- Rendering: Lighting models; Forward rendering; Inverse rendering; Image-based rendering; Non-photorealistic rendering; Graphics hardware; Visibility computation
- Shape: Surface reconstruction; Range imaging; Geometric modelling; Parameterization;
- Mesh simplification
- Animation: Motion capture and processing; Physics-based modelling; Character animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape Representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and Video Retrieval
- Video analysis and event recognition
- Medical Image Analysis
- Robot Localization and Navigation

Information for Librarians
Foundations and Trends® in Computer Graphics and Vision, 2006, Volume 2, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Foundations and Trends® in
Computer Graphics and Vision
Vol. 2, No. 4 (2006) 259–362

DOI: 10.1561/0600000018

now
the essence of knowledge

A Stochastic Grammar of Images

Song-Chun Zhu[1,*] and David Mumford[2]

[1] *University of California, Los Angeles, USA, sczhu@stat.ucla.edu*
[2] *Brown University, USA, David_Mumford@brown.edu*

Abstract

This exploratory paper quests for a stochastic and context sensitive grammar of images. The grammar should achieve the following four objectives and thus serves as a unified framework of representation, learning, and recognition for a large number of object categories. (i) The grammar represents both the hierarchical decompositions from scenes, to objects, parts, primitives and pixels by terminal and non-terminal nodes and the contexts for spatial and functional relations by horizontal links between the nodes. It formulates each object category as the set of all possible valid configurations produced by the grammar. (ii) The grammar is embodied in a simple And–Or graph representation where each Or-node points to alternative sub-configurations and an And-node is decomposed into a number of components. This representation supports recursive top-down/bottom-up procedures for image parsing under the Bayesian framework and make it convenient to scale up in complexity. Given an input image, the image parsing task constructs a most probable parse graph on-the-fly as the output interpretation and this parse graph is a subgraph of the And–Or graph after

[*] Song-Chun Zhu is also affiliated with the Lotus Hill Research Institute, China.

making choice on the Or-nodes. (iii) A probabilistic model is defined on this And–Or graph representation to account for the natural occurrence frequency of objects and parts as well as their relations. This model is learned from a relatively small training set per category and then sampled to synthesize a large number of configurations to cover novel object instances in the test set. This generalization capability is mostly missing in discriminative machine learning methods and can largely improve recognition performance in experiments. (iv) To fill the well-known semantic gap between symbols and raw signals, the grammar includes a series of visual dictionaries and organizes them through graph composition. At the bottom-level the dictionary is a set of image primitives each having a number of anchor points with open bonds to link with other primitives. These primitives can be combined to form larger and larger graph structures for parts and objects. The ambiguities in inferring local primitives shall be resolved through top-down computation using larger structures. Finally these primitives forms a primal sketch representation which will generate the input image with every pixels explained. The proposal grammar integrates three prominent representations in the literature: stochastic grammars for composition, Markov (or graphical) models for contexts, and sparse coding with primitives (wavelets). It also combines the structure-based and appearance based methods in the vision literature. Finally the paper presents three case studies to illustrate the proposed grammar.

Contents

1

Introduction

1.1 The Hibernation and Resurgence of Image Grammars

Understanding the contents of images has always been the core problem in computer vision with early work dated back to Fu [22], Riseman [33], Ohta and Kanade [54, 55] in the 1960–1970s. By analogy to natural language understanding, the task of image parsing [72], as Figure 1.1 illustrates, is to compute a parse graph as the most probable interpretation of an input image. This parse graph includes a tree structured decomposition for the contents of the scene, from scene labels, to objects, parts, primitives, so that all pixels are explained, and a number of spatial and functional relations between nodes for contexts at all levels of the hierarchy.

People who worked on image parsing in the 1960–1970s were, obviously, ahead of their time. In Kanade's own words, they had only 64K memory to work with at that time. Indeed, his paper with Ohta [55] was merely 4-page long! The image parsing efforts and structured methods encountered overwhelming difficulties in the 1970s and since then entered a hibernation state for a quarter of a century. The syntactic and grammar work have been mostly studied in the backstage as we

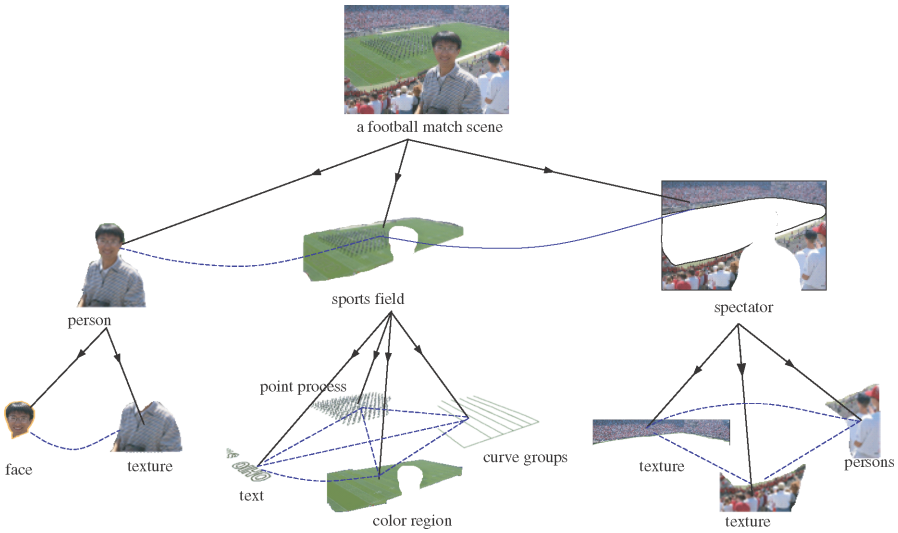

Fig. 1.1 Illustrating the task of image parsing. The parse graph includes a tree structured decomposition in vertical arrows and a number of spatial and functional relations in horizontal arrows. From [72].

shall review in later section. These difficulties remain challenging even today.

Problem 1: There is an enormous amount of visual knowledge about the real world scenes that has to be represented in the computer in order to make robust inference. For example, there are at least 3,000 object categories[1] and many categories have wide intra-category structural variations. The key questions are: how does one define an object category, say a car or a jacket? and how does one represent these categories in a consistent framework?

The visual knowledge is behind our vivid dreams and imaginations as well as the top-down computation. It was known that there are far more downward fibers than upward fibers in the visual pathways of primate animals. For example, it is reported in [65] that only 5%–10% of the input to the geniculate relay cells derives from the retina. The

[1] This number comes from Biederman who adopted a method used by pollsters. Take an English dictionary, open some pages at random, and count the number of nouns which are object categories at a page and then times the number of pages of the dictionary proportionally.

rest derives from local inhibitory inputs and descending inputs from layer 6 of the visual cortex. The weakness in knowledge representation and top-down inference is, in our opinion, the main obstacle in the road toward robust and large scale vision systems.

Problem 2: The computational complexity is huge.[2] A simple glance of Figure 1.1 reveals that an input image may contain a large number of objects. Human vision is known [70] to simultaneously activate the computation at all levels from scene classification to edge detection — all occurs in a very short time $\leq$400 ms, and to adopt multiple visual routines [76] to achieve robust computation. In contrast, most pattern recognition or machine learning algorithms are feedforward and computer vision systems rarely possess enough visual knowledge for reasoning.

The key questions are: how does one achieve robust computation that can be scaled to thousands of categories? and how does one coordinate these bottom-up and top-down procedures? To achieve scalable computation, the vision algorithm must be based on simple procedures and structures that are common to all categories.

Problem 3: The most obvious reason that sent the image parsing work to dormant status was the so-called semantic gap between the raw pixels and the symbolic token representation in early syntactic and structured methods. That is, one cannot reliably compute the symbols from raw images. This has motivated the shift of focus to appearance based methods in the past 20 years, such as PCA [75], AAM [12], and appearance based recognition [51], image pyramids [69] and wavelets [15], and machine learning methods [21, 63, 78] in the past decade.

Though the appearance based methods and machine learning algorithms have made remarkable progress, they have intrinsic problems that could be complemented by structure based methods. For example, they require too many training examples due to the lack the compositional and generative structures. They are often over-fit to specific training set and can hardly generalize to novel instances or configurations especially for categories that have large intra-class variations.

[2] The NP-completeness is no longer an appropriate measure of complexity, because even many simplified vision problems are known to be NP-hard.

After all these developments, the recent vision literature has observed a pleasing trend for returning to the grammatical and compositional methods, for example, the work in the groups of Ahuja [71], Geman [27, 36], Dickinson [14, 40], Pollak [79], Buhmann [57] and Zhu [9, 32, 44, 59, 72, 74, 85, 86]. The return of grammar is in response to the limitations of the appearance based and machine learning methods when they are scaled up.

The return of grammar is powered by progresses in several aspects, which were not available in the 1970s. (i) A consistent mathematical and statistical framework to integrate various image models, such as Markov (graphical) models [90], sparse coding [56], and stochastic context free grammar [10]. (ii) More realistic appearance models for the image primitives to connect the symbols to pixels. (iii) More powerful algorithms including discriminative classification and generative methods, such as the Data-Driven Markov China Monte Carlo (DDMCMC) [73]. (iv) Huge number of realistic training and testing images [87].

1.2 Objectives

This exploratory paper will review the issues and recent progress in developing image grammars, and introduce a stochastic and context sensitive grammar as a unified framework for representation, learning, and recognition. This framework integrates many existing models and algorithms in the literature and addresses the problems raised in the previous subsection. This image grammar should achieve the following four objectives.

Objective 1: *A common framework for visual knowledge representation and object categorization.* Grammars, studied mostly in language [1, 26], are known for their expressive power in generating a very large set of configurations or instances, i.e., their language, by composing a relatively much smaller set of words, i.e., shared and reusable elements, using production rules. Hierarchic and structural composition is the key concept behind grammars in contrast to enumerating all possible configurations.

In this paper, we embody the image grammar in an And–Or graph representation[3] where each Or-node points to alternative sub-configurations and an And-node is decomposed into a number of components. This And–Or graph represents both the hierarchical decompositions from scenes, to objects, parts, primitives and pixels by terminal and non-terminal nodes and the contexts for spatial and functional relations by horizontal links between the nodes. It is an alternate way of representing production rules and it contains all possible parse trees. Then we will define a probabilistic model for the And–Or graph which can be learned from examples using maximum likelihood estimation. Therefore, all the structural and contextual information are represented in the And–Or graph (and equivalently the grammar). This also resolve the object categorization problem. We can define each object category as *the set of all valid configurations which are produced by the grammar, with its probability learned to reproduce natural frequency of instances occurring in the observed ensemble.*

As we will show in later section, this probability model integrates popular generative models, such as sparse coding (wavelet coding) and stochastic context free grammars (SCFG), with descriptive models, such as Markov random fields and graphical models. The former represents the generative hierarchy for reconfigurability while the latter models context.

Objective 2: *Scalable and recursive top-down/bottom-up computation.* The And–Or graph representation has recursive structures with two types of nodes. It can be easily scalable up in the number of nodes and object categories. For example, suppose an Or-node represents an object, say car, it then has a number of children nodes for different views (front, side, back etc.) of cars. By adding a new child node, we can augment to new views. This representation supports recursive top-down/bottom-up procedures for image parsing and make it convenient to scale up in complexity.

Figure 1.2 shows a parsing graph under construction at a time step. This simple grammar is one of our case study in later section uses one

[3] The And–Or graph was previously used by Pearl in [58] for heuristic searches. In our work, we use it in a very different purpose and should not be confused with Pearl's work.

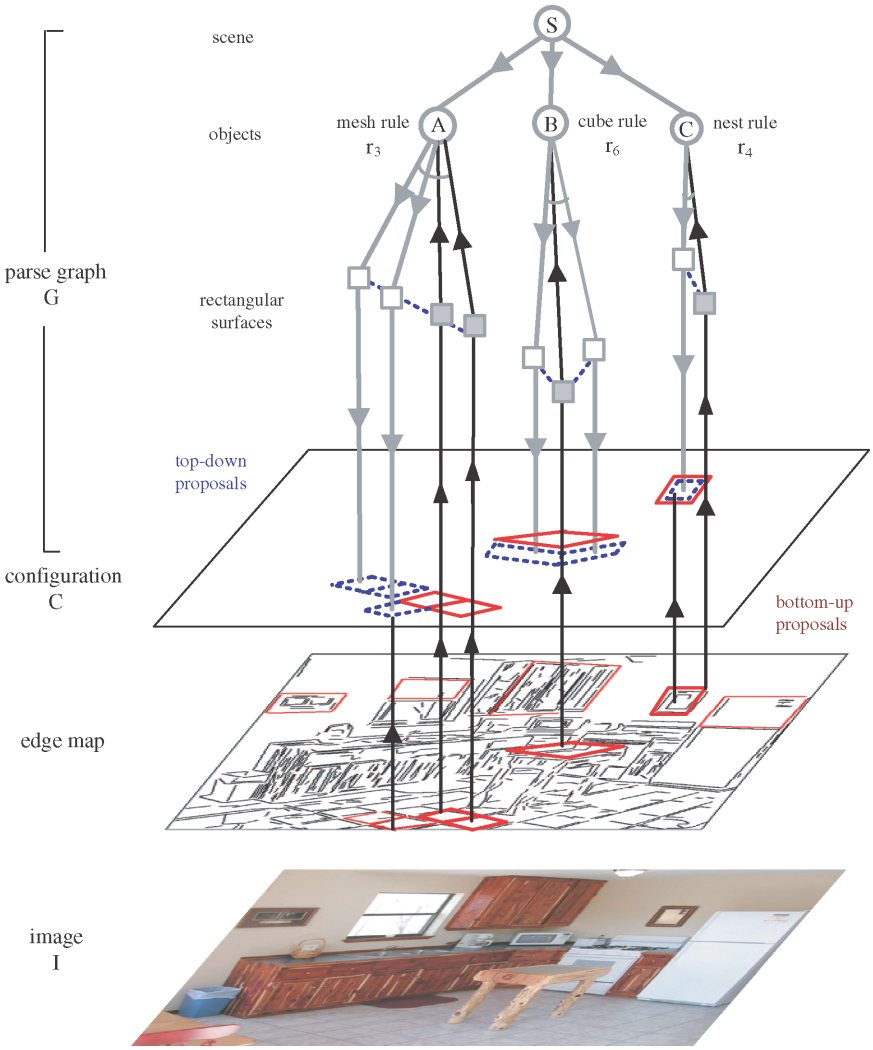

Fig. 1.2 Illustrating the recursive bottom-up/top-down computation processes in image parsing. The detection of rectangles (in red) instantiates some non-terminal nodes shown as upward arrows. They in turn activate graph grammar rules for grouping larger structures in nodes A, B, and C, respectively. These rules generate top-down prediction of rectangles (in blue). The predictions are validated from the image under the Bayesian posterior probability. Modified from [59].

primitive: rectangular surfaces projected onto the image plane. The grammar rules represents various organization, such as alignments of the rectangles in mesh, linear, nesting, cubic structures. In the kitchen scene, the four rectangles (in red) accepted through bottom-up process and they activate the production rules represented by the non-terminal nodes A, B, and C, respectively. Which then predict a number of candidates (in blue) in top-down search. The solid upward arrows show the bottom-up binding, while the downward arrows show the top-down prediction. As the ROC curves in Figure 9.5 shows in later section, the top-down prediction largely improves the recognition rate of the rectangles, as certain rectangles can only be hallucinated through top-down process due to occlusion and severe image degradation.

Given an input image, the image parsing task constructs a most probable parse graph on-the-fly as the output interpretation and this parse graph is a subgraph of the And–Or graph after making choices on the Or-nodes.

As we shall discuss in later section, the computational algorithm maintains the same data structures for each of the And-nodes and Or-nodes in the And–Or graph and adopt the same computational procedure: (i) bottom-up detecting and binding using a cascade of features; and (ii) top-down on-line template composition and matching. To implement the system, we only need to write one common class (in C++ programming) for all the nodes, and different objects and parts are realized as instances of this class. These nodes use different bottom-up features/tests and the top-down templates during the computational process. The features and templates are learned off-line through training images and loaded into the instances of the C++ class during the computational process. This recursive algorithm has the potential to be implemented in a massively parallel machine where each unit has the same data structures and functions described above.

Objective 3: *Small sample learning and generalization.* The probabilistic model defined on this And–Or graph representation can be learned from a relatively small training set per category and then sampled through Monte Carlo simulation to synthesize a large number of configurations. This is in fact an extension to the traditional texture synthesis experiment by the minimax entropy principle [90], where new

texture samples are synthesized which are different from the observed texture but are perceptually equivalent to the observed texture. The minimax entropy learning scheme is extended to the And–Or graph models in [59], which can generate novel configurations through composition to cover unforeseen object instances in the test set. This generalization capability is mostly missed in discriminative machine learning methods.

In the experiments reported in [44, 59], they seek for the minimum number of distinct training samples needed for each category, usually in the range of 20–50. They prune some redundant examples which can be derived through other examples by composition. Then they found that the generated samples can largely improve the object recognition performance. For example, a 15% recognition rate is reported in [44].

Objective 4: *Mapping the visual vocabulary to fill the semantic gap.* To fill the well-known semantic gap between symbols and pixels, the grammar includes a series of visual dictionaries for visual concepts at all levels. There are two key observations for these dictionaries.

1. The elements of the dictionaries are organized through graph composition. At the bottom-level the dictionary is a set of image primitives each having a number of anchor points in a small graph with open bonds to link with other primitives. These primitives can be combined to form larger and larger graph structures for parts and objects, in a way similar to Lego pieces that kids play with.[4]
2. Vision is distinct from other sensors, like speech in the aspect that objects can appear at arbitrary scales. As a result, the instances of each node can occur at any sizes. The non-terminal nodes at all levels of the And–Or graph can terminate directly as image primitives. Thus one has to account for the transitions between instances of the same node over scales. This is the topics studied in the perceptual scale space theory [80].

[4] Note that Lego pieces are well designed to have standardized teeth to fit each other, this is not true in the image primitives. The latter are more flexible.

Though there are variations in the literature for what the low level primitives should be, the differences are really minor between what people called textons, texels, primitives, patches, and fragments. The ambiguities in inferring these local primitives shall be resolved through top-down computation using larger structures.

Finally the primitives are connected to form a primal sketch graph representation [31] which will generate the input image with every pixels explained. This closes the semantic gap.

1.3 Overview of the Image Grammar

In this subsection, we overview the basic concepts in the image grammar. We divided it into two parts: (i) representation and data structures, (ii) Image annotation dataset to learn the grammar, and the learning and computing issues.

1.3.1 Overview of the Representational Concepts and Data Structures

We use Figure 1.3 as an example to review the representational concepts in the following:

1. *An And–Or graph.* Figure 1.3(a) shows a simple example of an And–Or graph. An And–Or graph includes three types of nodes: And-nodes (solid circles), Or-nodes (dashed circles), and terminal nodes (squares). An And-node represents a decomposition of an entity into its parts. It corresponds to the grammar rules, for example,

$$A \rightarrow BCD, \quad H \rightarrow NO.$$

The horizontal links between the children of an And-node represent relations and constraints. The Or-nodes act as "switches" for alternative sub-structures, and stands for labels of classification at various levels, such as scene category, object classes, and parts etc. It corresponds to production rules like,

$$B \rightarrow E \mid F, \quad C \rightarrow G \mid H \mid I.$$

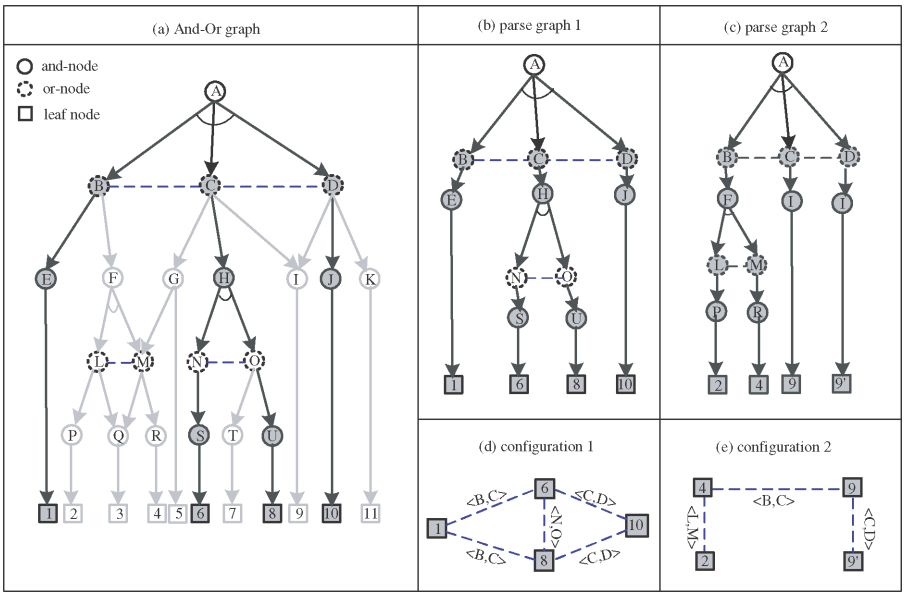

Fig. 1.3 Illustrating the And–Or graph representation. (a) An And–Or graph embodies the grammar productions rules and contexts. It contains many parse graphs, one of which is shown in bold arrows. (b) and (c) are two distinct parse graphs by selecting the switches at related Or-nodes. (d) and (e) are two graphical configurations produced by the two parse graphs, respectively. The links of these configurations are inherited from the And–Or graph relations. Modified from [59].

Due to this recursive definition, one may merge the And–Or graphs for many objects or scene categories into a larger graph. In theory, all scene and object categories can be represented by one huge And–Or graph, as it is the case for natural language. The nodes in an And–Or graph may share common parts, for example, both cars and trucks have rubber wheels as parts, and both clock and pictures have frames.

2. *A parse graph*, as shown in Figure 1.1, is a hierarchic generative interpretation of a specific image. A parse graph is augmented from a parse tree, mostly used in natural or programming language by adding a number of relations, shown as side links, among the nodes. A parse graph is derived from the And–Or graph by selecting the switches or classification labels at related Or-nodes. Figures 1.3(b) and 1.3(c)

are two instances of the parse graph from the And–Or graph in Figure 1.3(a). The part shared by two node may have different instances, for example, node I is a child of both nodes C and D. Thus we have two instances for node 9.

3. *A configuration* is a planar attribute graph formed by linking the open bonds of the primitives in the image plane. Figures 1.3(d) and 1.3(e) are two configurations produced by the parse graphs in Figures 1.3(b) and 1.3(c), respectively. Intuitively, when the parse graph collapses, it produces a planar configuration. A configuration inherits the relations from its ancestor nodes, and can be viewed as a Markov networks (or deformable templates [19]) with reconfigurable neighborhood. We introduce a mixed random field model [20] to represent the configurations. The mixed random field extends conventional Markov random field models by allowing address variables and handles non-local connections caused by occlusions. In this generative model, a configuration corresponds to a primal sketch graph [31].
4. *The visual vocabulary.* Due to scaling property, the terminal nodes could appear at all levels of the And–Or graph. Each terminal node takes instances from certain set. The set is called a dictionary and contains image patches of various complexities. The elements in the set may be indexed by variables such as its type, geometric transformations, deformations, appearance changes etc. Each patch is augmented with anchor points and open bond to connect with other patches.
5. *The language* of a grammar is the set of all possible valid configurations produced by the grammar. In stochastic grammar, each configuration is associated with a probability. As the And–Or graph is directed and recursive, the sub-graph underneath any node A can be considered a sub-grammar for the concept represented by node A. Thus a sub-language for node A is the set of all valid configurations produced by the And–Or graph rooted at A. For example, if A is an object category, say a car, then this sub-language defines all the valid

> configurations of car. In an exiting case, the sub-language of a terminal node contains only the atomic configurations and thus is called a dictionary.

In comparison, an element in a dictionary is an atomic structure and an element in a language is a composite structure (or configuration) made of a number of atomic structures. A configuration of node A in zoomed-out view loses its resolution and details, and becomes an atomic element in the dictionary of node A. For example, a car viewed in close distance is a configuration consisting of many parts and primitives. But in far distance, a car is represented by a small image patch as a whole and is not decomposable. This is a special property of the image grammar. The perceptual transition over scales is studied in [80, 84].

1.3.2 Overview of the Dataset and Learning

Now we briefly overview the learning and computing issues with stochastic image grammars.

A foremost question that one may ask is: how do you build this grammar and where is the dataset? Collecting the dataset for learning and training is perhaps more challenging than the learning task itself.

Although fully automated learning is most ideal, for example, let a computer program watch Disney cartoon or Hollywood movies and hope it figures out all the object categories and relations. But purely unsupervised learning is less practical for learning the structured compositional models at present for two reasons. (i) Visual learning must be guided by objectives and purposes of vision, not purely based on statistical information. Ideally one has to integrate this automatic learning process with autonomous robot and AI reasoning at the higher level. Before the robotics and AI systems are ready, we should guide the learning process with some human supervision. For example, what are important structures and what are decorative stuff. (ii) In almost all the unsupervised learning methods, the trainers still have to select their data carefully to contrast the involved concepts. For example, to learn the concept that a car has doors, we must select images of cars with doors both open and closed. Otherwise the concept of door cannot be learned.

We propose to learn the image grammar in a semi-automatic way. We shall start with a supervised learning with manually annotated images and objects to produce the parse graphs. We use this dataset to initiate the process and then shift to weakly supervised learning. This initial dataset is still very large if we target thousands of object categories.

To make the large scale grammar learning framework practical, the first author founded an independent non-profit research institute which started to operate in the summer of 2005.[5] It has a full time annotation team for parsing the image structures and a development team for the annotation tools and database construction. Each image or object is parsed, semi-automatically, into a *parse graph* where the relations are specified and objects are names using the wordnet standard. Figure 1.4 lists an inventory of the current ground truth dataset parsed at LHI. It has now over 500,000 images (or video frames) parsed, covering 280 object categories. Figure 1.5 shows two examples — the parse trees of cat and car. For clarity we only show the parse trees with naming of the nodes. Beyond the object parsing, there are many scene images annotated with the objects and their spatial relations labeled. As stated in a report [87], this ground truth annotation is aimed at broader scope and more hierarchic structures than other datasets collected in various groups, such as Berkeley [4, 50], Caltech [16, 29], and MIT [62].

With this annotated dataset, we can construct the And–Or graph for object and scene categories and learn the probability model on the And–Or graphs. These learning steps are guided by a minimax entropy learning scheme [90] and maximum likelihood estimation. It is divided into three parts:

1. Learning the probabilities at the Or-node so that the configurations generated account for the natural co-occurrence frequency. This is typical in stochastic context free grammars [10].
2. Learning and pursuing the Markov models on the horizontal links and relations to account for the spatial relations, as well

[5] It is called the Lotus Hill Research Institute (LHI) in China (www.lotushill.org).

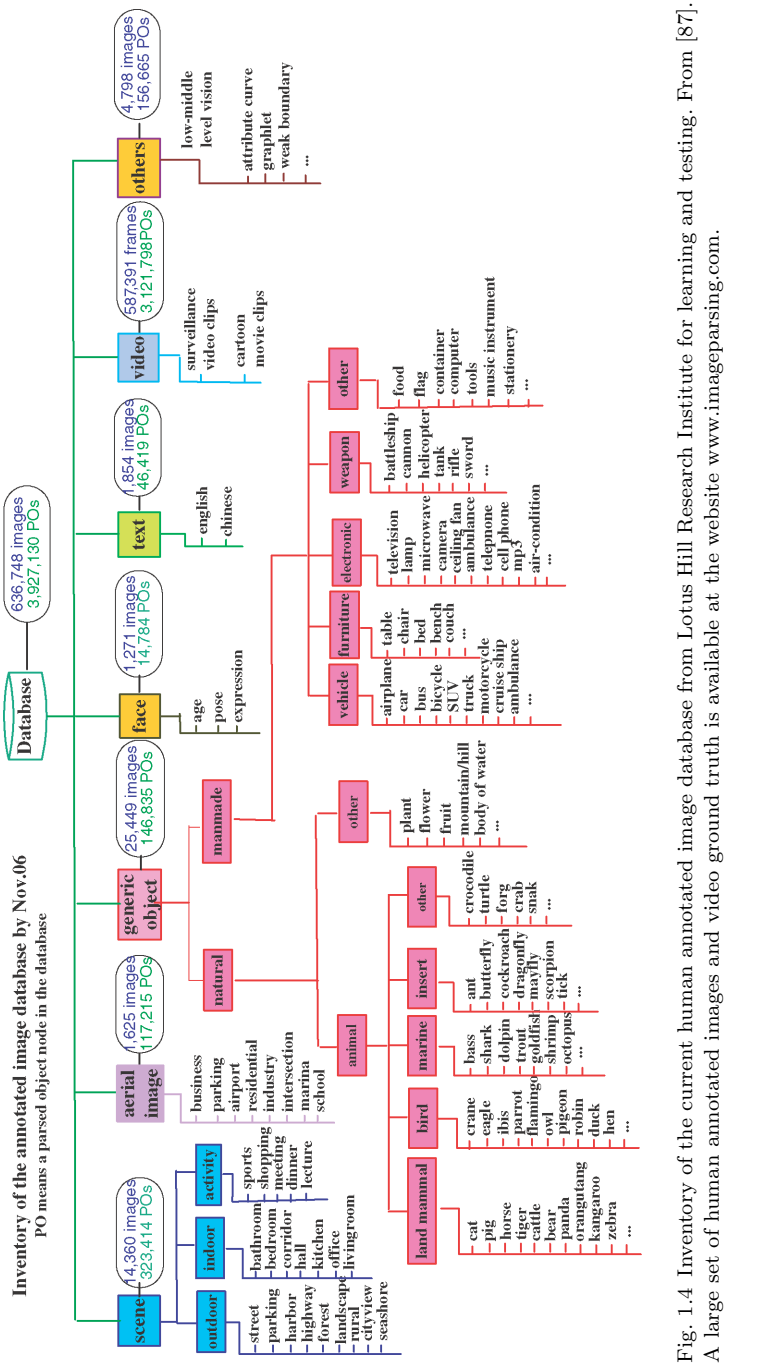

Fig. 1.4 Inventory of the current human annotated image database from Lotus Hill Research Institute for learning and testing. From [87]. A large set of human annotated images and video ground truth is available at the website www.imageparsing.com.

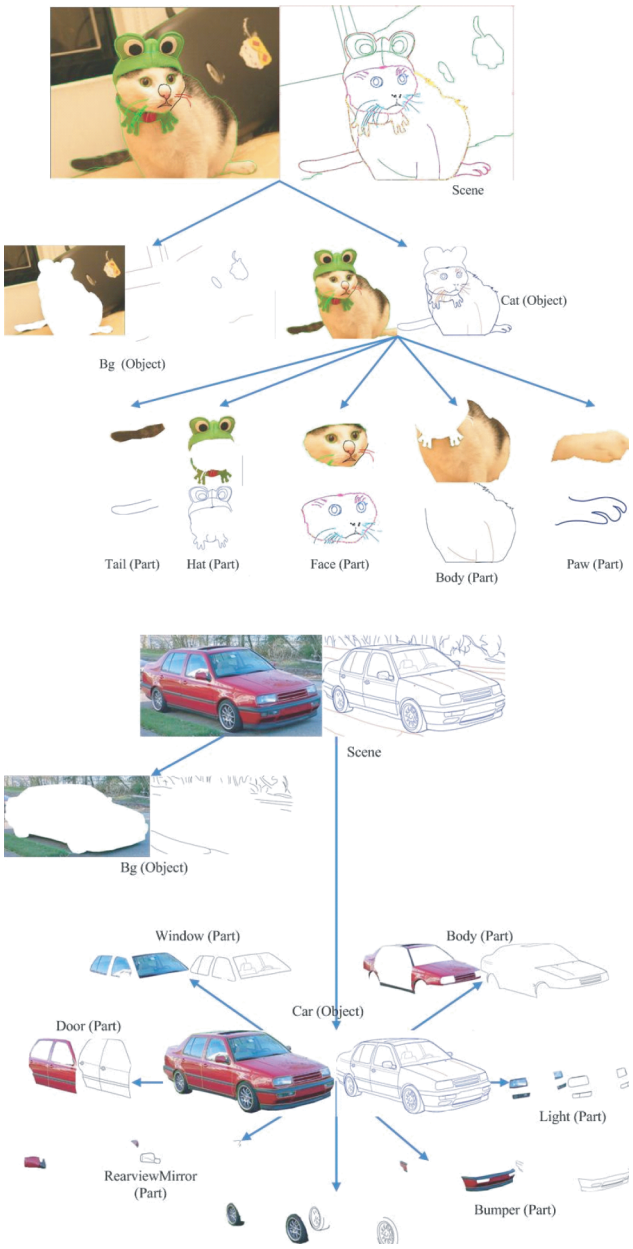

Fig. 1.5 Two examples of the parse trees (cat and car) in the Lotus Hill Research Institute image corpus. From [87].

as consistency of appearance between nodes in the And–Or graphs. This is similar to the learning of Markov random fields [90], except that we are dealing with a dynamic graphical configuration instead of a fixed neighborhood.

3. Learning the And–Or graph structures and dictionaries. The terminal nodes are learned through clustering and the non-terminal nodes are learned through binding. We only briefly discuss this issue in this paper as the current literature has not made significant progress in this part.

The proposed stochastic context sensitive grammar (SCSG) combines the reconfigurability of SCFG with the contextual constraints of graphical (MRF) models, and has the following properties: (a) Compositional power for representing large intra-class structural variations. The grammar can generate a huge number of configurations (i.e., its language) for scenes and objects by composing a relatively much smaller vocabulary. All are represented in graphical configurations. The language of the grammar is the set of all valid configurations of a category, such as furniture, clothes, vehicles, etc. Thus it has enormous expressive power. (b) Recursive structures for scalable computing. The grammar is embodied into an And–Or graph which has recursive structure. The latter is easy to scale in terms of increasing the number of object categories or augmenting more levels (e.g., scene nodes). Consequently the inference algorithms is also recursively defined. We only need to write general top-down and bottom-up functions for a common And–Or node, and re-use the code for all nodes in the And–Or graph. (c) Small sample for effective learning. Due to explicit composition and part-sharing between categories, the state spaces for all object categories are decomposed into products of subspaces of lower dimensions for the vocabulary and relations. Thus we need relatively smaller number of training examples (20–100 instances) for each category. In recent experiments (see Figure 2.6), we can sample the learned object model to generate novel object configurations for generalization, and observe remarkable (over 15% improvement in object category) recognition tasks.

The rest of the paper is organized in the following way. We first discuss in Chapter 2 the background of stochastic grammar, its formulation, the new issues of image grammar in contrast to language grammar, and previous work on image grammar. Then we present the grammar and And–Or graph representation in Chapters 3–6 sequentially: the visual grammar, the relations and configurations, the parse graphs, and finally the And–Or graph. The learning algorithm and results are discussed in Chapter 7, which is followed by the top-down/bottom-up inference algorithm in Chapter 8, and three case studies in Chapter 9. Finally, we raise a number of unsolved problems in Chapter 10 to conclude the paper.

References

[1] S. P. Abney, "Stochastic attribute-value grammars," *Computational Linguistics*, vol. 23, no. 4, pp. 597–618, 1997.

[2] K. Athreya and A.Vidyashankar, *Branching Processes*. Springer-Verlag, 1972.

[3] A. Barbu and S. C. Zhu, "Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities," *IEEE Transactions on PAMI*, vol. 27, no. 8, pp. 1239–1253, 2005.

[4] K. Barnard *et al.*, "Evaluation of localized semantics: Data methodology, and experiments," Tech. Report, CS, U. Arizona, 2005.

[5] I. Biederman, "Recognition-by-components: A theory of human image understanding," *Psychological Review*, vol. 94, pp. 115–147, 1987.

[6] E. Bienenstock, S. Geman, and D. Potter, "Compositionality, MDL priors, and object Recognition," in *Advances in Neural Information Processing Systems 9*, (M. Mozer, M. Jordan, and T. Petsche, eds.), MIT Press, 1998.

[7] G. Blanchard and D. Geman, "Sequential testing designs for pattern recognition," *Annals of Statistics*, vol. 33, pp. 1155–1202, June 2005.

[8] H. Blum, "Biological shape and visual science," *Journal of Theoretical Biology*, vol. 38, pp. 207–285, 1973.

[9] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu, "Composite templates for cloth modeling and sketching," in *Proceedings of IEEE Conference on Pattern Recognition and Computer Vision*, New York, June 2006.

[10] Z. Y. Chi and S. Geman, "Estimation of probabilistic context free grammar," *Computational Linguistics*, vol. 24, no. 2, pp. 299–305, 1998.

[11] N. Chomsky, *Syntactic Structures*. Mouton: The Hague, 1957.

[12] T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham, "Active appearance models–their training and applications," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.
[13] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet based statistical signal processing using hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 46, pp. 886–902, 1998.
[14] S. J. Dickinson, A. P. Pentland, and A. Rosenfeld, "From volumes to views: An approach to 3D object recognition," *CVGIP: Image Understanding*, vol. 55, no. 2, pp. 130–154, 1992.
[15] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechie, "Data compression and harmonic analysis," *IEEE Transactions on Information Theory*, vol. 6, pp. 2435–2476, 1998.
[16] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 100 object categories," *Workshop on Generative Model Based Vision*, 2004.
[17] L. Fei-Fei, R. Fergus, and P. Perona, "One-Shot learning of object categories," *IEEE Transactions on PAMI*, vol. 28, no. 4, pp. 594–611, 2006.
[18] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?," *Journal of Vision*, vol. 7, no. 1, pp. 1–29, 2007.
[19] M. Fischler and R. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computer*, vol. C-22, pp. 67–92, 1973.
[20] A. Fridman, "Mixed markov models," *Proceedings of Natural Academy of Science USA*, vol. 100, pp. 8092–8096, 2003.
[21] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Annals of Statistics*, vol. 38, no. 2, pp. 337–374, 2000.
[22] K. S. Fu, *Syntactic Pattern Recognition and Applications*. Prentice-Hall, 1982.
[23] M. Galun, E. Sharon, R. Basri, and A. Brandt, "Texture segmentation by multiscale aggregation of filter responses and shape elements," *Proceedings of ICCV, Nice*, pp. 716–723, 2003.
[24] R. X. Gao, T. F. Wu, N. Sang, and S. C. Zhu, "Bayesian inference for layered representation with mixed Markov random field," in *Proceedings of the 6th International Conference on EMMCVPR*, Ezhou, China, August 2007.
[25] R. X. Gao and S. C. Zhu, "From primal sketch to 2.1D sketch," Technical Report, Lotus Hill Institute, 2006.
[26] S. Geman and M. Johnson, "Probability and statistics in computational linguistics, a brief review," in *Int'l Encyc. of the Social and Behavioral Sciences*, (N. J. Smelser and P. B. Baltes, eds.), pp. 12075–12082, Pergamon: Oxford, 2002.
[27] S. Geman, D. Potter, and Z. Chi, "Composition systems," *Quarterly of Applied Mathematics*, vol. 60, pp. 707–736, 2002.
[28] U. Grenander, *General Pattern Theory*. Oxford University Press, 1993.
[29] G. Griffin, A. Holub, and P. Perona, "The Caltech 256," Technical Report, 2006.
[30] C. E. Guo, S. C. Zhu, and Y. N. Wu, "Modeling visual patterns by integrating descriptive and generative models," *IJCV*, vol. 53, no. 1, pp. 5–29, 2003.

[31] C. E. Guo, S. C. Zhu, and Y. N. Wu, "Primal sketch: Integrating texture and structure," in *Proceedings of International Conference on Computer Vision*, 2003.
[32] F. Han and S. C. Zhu, "Bottom-up/top-down image parsing by attribute graph grammar". *Proceedings of International Conference on Computer Vision,* Beijing, China, 2005. (A long version is under review by PAMI).
[33] A. Hanson and E. Riseman, "Visions: A computer system for interpreting scenes," in *Computer Vision Systems*, 1978.
[34] T. Hong and A. Rosenfeld, "Compact region extraction using weighted pixel linking in a pyramid," *IEEE Transactions on PAMI*, vol. 6, pp. 222–229, 1984.
[35] J. Huang, PhD Thesis, Division of Applied Math, Brown University.
[36] Y. Jin and S. Geman, "Context and hierarchy in a probabilistic image model," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, New York, June 2006.
[37] B. Julesz, "Textons, the elements of eexture perception, and their interactions," *Nature*, vol. 290, pp. 91–97, 1981.
[38] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, 2001.
[39] G. Kanisza, *Organization in Vision*. New York: Praeger, 1979.
[40] Y. Keselman and S. Dickinson, "Generic model abstraction from examples," *CVPR*, 2001.
[41] B. Kimia, A. Tannenbaum, and S. Zucker, "Shapes, shocks and deformations I," *Interantional Journal of Computer Vision*, vol. 15, pp. 189–224, 1995.
[42] A. B. Lee, K. S. Pedersen, and D. Mumford, "The nonlinear statistics of high-contrast patches in natural images," *IJCV*, vol. 54, no. 1/2, pp. 83–103, 2003.
[43] M. Leyton, "A process grammar for shape," *Artificial Intelligence*, vol. 34, pp. 213–247, 1988.
[44] L. Lin, S. W. Peng, and S. C. Zhu, "An empirical study of object category recognition: Sequential testing with generalized samples," in *Proceedings of International Conference on Computer Vision*, Rio de Janeiro, Brazil, October 2007.
[45] T. Lindeberg, *Scale-Space Theory in Computer Vision*. Netherlands: Kluwer Academic Publishers, 1994.
[46] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. NY: Springer-Verlag, p. 134, 2001.
[47] S. Mallat and Z. Zhang, "Matching pursuit in a time-frequency dictionary," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.
[48] K. Mark, M. Miller, and U. Grenander, "Constrained stochastic language models," in *Image Models (and Their Speech Model cousins)*, (S. Levinson and L. Shepp, eds.), IMA Volumes in Mathematics and its Applications, 1994.
[49] D. Marr, *Vision*. Freeman Publisher, 1983.
[50] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms," *ICCV*, 2001.

[51] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-D objects from appearance," *International Journal of Computer Vision*, vol. 14, pp. 5–24, 1995.

[52] K. Murphy, A. Torralba, and W. T. Freeman, "Graphical model for recognizing scenes and objects," *Proceedings of NIPS*, 2003.

[53] M. Nitzberg, D. Mumford, and T. Shiota, "Filtering, segmentation and depth," *Springer Lecture Notes in Computer Science*, vol. 662, 1993.

[54] Y. Ohta, *Knowledge-Based Interpretation of Outdoor Natural Color Scenes*. Pitman, 1985.

[55] Y. Ohta, T. Kanade, and T. Sakai, "An analysis system for scenes containing objects with substructures," in *Proceedings of 4th International Joint Conference on Pattern Recognition*, (Kyoto), pp. 752–754, 1978.

[56] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[57] B. Ommer and J. M. Buhmann, "Learning compositional categorization method," in *Proceedings of European Conference on Computer Vision*, 2006.

[58] J. Pearl, *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, 1984.

[59] J. Porway, Z. Y. Yao, and S. C. Zhu, "Learning an And–Or graph for modeling and recognizing object categories," Technical Report, Department of Statistics, UCLA, 2007.

[60] J. Rekers and A. Schürr, "A parsing algorithm for context sensitive graph grammars," TR-95–05, Leiden University, 1995.

[61] M. Riesenhuber and T. Poggio, "Neural mechanisms of object recognition," *Current Opinion in Neurobiology*, vol. 12, pp. 162–168, 2002.

[62] B. Russel, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A database and web-based tool for image annotation," *MIT AI Lab Memo AIM-2005-025*, September 2005.

[63] R. E. Schapire, "The boosting approach to machine learning: An overview," *MSRI Workshop on nonlinear Estimation and Classification*, 2002.

[64] T. B. Sebastian, P. N. Klein, and B. B. Kimia, "Recognition of shapes by editing their shock graphs," *IEEE Transactions on PAMI*, vol. 26, no. 5, pp. 550–571, 2004.

[65] S. M. Sherman and R. W. Guillery, "The role of thalamus in the flow of information to cortex," *Philosophical Transactions of Royal Society London (Biology)*, vol. 357, pp. 1695–1708, 2002.

[66] K. Shi and S. C. Zhu, "Visual learning with implicit and explicit manifolds," *IEEE Conference on CVPR*, June 2007.

[67] K. Siddiqi and B. B. Kimia, "Parts of visual form: Computational aspects," *IEEE Transactions on PAMI*, vol. 17, no. 3, pp. 239–251, 1995.

[68] K. Siddiqi, A. Shokoufandeh, S. J. Dickinson, and S. W. Zucker, "Shock graphs and shape matching," *IJCV*, vol. 35, no. 1, pp. 13–32, 1999.

[69] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 587–607, 1992.

[70] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520–522, 1996.
[71] S. Todorovic and N. Ahuja, "Extracting subimages of an unknown category from a set of images," *CVPR*, 2006.
[72] Z. W. Tu, X. R. Chen, A. L. Yuille, and S. C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 113–140, 2005.
[73] Z. W. Tu and S. C. Zhu, "Image segmentation by data-driven Markov chain Monte Carlo," *IEEE Transactions on PAMI*, May 2002.
[74] Z. W. Tu and S. C. Zhu, "Parsing images into regions, curves and curve groups," *International Journal of Computer Vision*, vol. 69, no. 2, pp. 223–249, 2006.
[75] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, p. 1, 1991.
[76] S. Ullman, "Visual routine," *Cognition*, vol. 18, pp. 97–157, 1984.
[77] S. Ullman, E. Sali, and M. Vidal-Naquet, "A fragment-based approach to object representation and classification," in *Proceedings of 4th International Workshop on Visual Form*, Capri, Italy, 2001.
[78] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *CVPR*, pp. 511–518, 2001.
[79] W. Wang, I. Pollak, T.-S. Wong, C. A. Bouman, M. P. Harper, and J. M. Siskind, "Hierarchical stochastic image grammars for classification and segmentation," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3033–3052, 2006.
[80] Y. Z. Wang, S. Bahrami, and S. C. Zhu, "Perceptual scale space and it applications," in *International Conference on Computer Vision*, Beijing, China, 2005.
[81] M. Weber, M. Welling, and P. Perona, "Towards automatic discovery of object categories," *IEEE Conference on CVPR*, 2000.
[82] A. P. Witkin, "Scale space filtering," *International Joint Conference on AI*. Palo Alto: Kaufman, 1983.
[83] T. F. Wu, G. S. Xia, and S. C. Zhu, "Compositional boosting for computing hierarchical image structures," *IEEE Conference on CVPR*, June 2007.
[84] Y. N. Wu, S. C. Zhu, and C. E. Guo, "From information scaling laws of natural images to regimes of statistical models," *Quarterly of Applied Mathematics*, 2007 (To appear).
[85] Z. J. Xu, H. Chen, and S. C. Zhu, "A high resolution grammatical model for face representation and sketching," in *Proceedings of IEEE Conference on CVPR*, San Diego, June 2005.
[86] Z. J. Xu, L. Lin, T. F. Wu, and S. C. Zhu, "Recursive top-down/bottom-up algorithm for object recognition," Technical Report, Lotus Hill Research Institute, 2007.
[87] Z. Y. Yao, X. Yang, and S. C. Zhu, "Introduction to a large scale general purpose groundtruth database: Methodology, annotation tools, and benchmarks," in *6th International Conference on EMMCVPR*, Ezhou, China, 2007.
[88] S. C. Zhu, "Embedding Gestalt laws in Markov random fields," *IEEE Transactions on PAMI*, vol. 21, no. 11, 1999.

[89] S. C. Zhu, "Statistical modeling and conceptualization of visual patterns," *IEEE Transactions on PAMI*, vol. 25, no. 6, pp. 691–712, 2003.

[90] S. C. Zhu, Y. N. Wu, and D. B. Mumford, "Minimax entropy principle and its applications to texture modeling," *Neural Computation*, vol. 9, no. 8, pp. 1627–1660, November 1997.

[91] S. C. Zhu and A. L. Yuille, "Forms: A flexible object recognition and modeling system," *Interantional Journal of Computer Vision*, vol. 20, pp. 187–212, 1996.

[92] S. C. Zhu, R. Zhang, and Z. W. Tu, "Integrating top-down/bottom-up for object recognition by data-driven Markov chain Monte Carlo," *CVPR*, 2000.