Natural Scene Segmentation of Static Images

Erich Leung

John Tsotsos

Technical Report CSE-2009-03

March 28 2009

Department of Computer Science and Engineering

4700 Keele Street Toronto, Ontario M3J 1P3 Canada

Abstract

This paper seeks to provide a systematic account of the conceptual and computational principles pertinent to recent approaches to natural scene segmentation. Image segmentation has been traditionally viewed as a problem of partitioning an image into regions with respect to content-independent visual structures – a class of image segmentation referred to in this paper as image-based segmentation. Also developed are different approaches to content-specific segmentation of natural scenes with respect to meaningful visual entities – the approaches collectively referred to as semantic-based natural scene segmentation. The main interest of the paper is concerned with this latter class of image segmentation. In particular, its focus is drawn to the recent approaches which apply probabilistic methods of visual classification as well as image segmentation methodologies to decomposing natural scenes with respect to object classes. This class of semantic-based image segmentation is referred to as probabilistic approaches to natural scene segmentation.

The paper starts the account by highlighting a common set of conceptual underpinnings of image segmentation across natural scene segmentation methodologies. In particular, Section 2 lays bare the nature and the challenges of image-based natural image segmentation and describes the recent shift of attention in a number of approaches to the semantic aspects of the problem. The rest of the paper is concerned with the major ideas and techniques of probabilistic approaches. Section 2.3 describes the problems as well as the major ideas and conceptual princi-

ples of probabilistic approaches. The section also outlines the two major approaches to representing the inference problems of natural scene segmentation, viz, image partition via stochastic search and scene labeling via visual classification. Section 4 discusses the conceptual issues pertaining to (1) how to capture semantics from visual features, (2) how to integrate image-based and semantic-based visual cues, and (3) how to organize relevant information into a coherent system of representation and inference. The last section focuses on the conceptual and empirical issues of the prevailing models in the current developments as well as the open issues that worth attention for future research. Supplementary materials are provided in the appendix, which summarizes some major probabilistic approaches to natural scene segmentation.

Contents

1 Introduction

1.1 Scope of Discussion

The goal of image segmentation is to determine the parts of an image that admit being perceived as a meaningful whole or segments [3]. A subimage – a collection of observed units, referred to in this paper as visual inputs, such as pixels or groups of pixels – is interpreted as a segment if the input can be explained in terms of a set of physical, geometric as well as semantic properties or features. These pieces are expected to yield a meaningful decomposition of sensory input, that represents a coherent, semantic interpretation of the external environment. A good image segmentation can be beneficial in a wide range of computational visual tasks, including high quality image/video editing, object-based encoding, transmission and manipulation of multimedia data, context-based image analysis, object recognition and others. [33, 97].

Over decades of active research, a vast literature has developed, characterized by a proliferation of frameworks built on a diversity of theoretical, mathematical and algorithmic underpinnings and the cross-fertilization between these ideas. Any attempt at a non-trivial overview can easily grow into a book-length study. As a result, this paper restricts its scope to some emerging paradigms with a focus on some previously less explored problems of image segmentation – in particular, the issues concerning the role of semantic content of visual input and their delineation.

The challenge of natural scene segmentation is great. A thoughtful examina-

tion of this class of problems may help to highlight what image segmentation may entail in general. For this purpose, the term 'natural images' refers to any image of a non-constrained and non-contrived scene which may consist of objects, both man-made and otherwise, in some indoor, outdoor or combined conditions[1]. It may be natural for biological observers to perceive their environments dynamically. For many reasons, however, the main focus of this paper is segmentation of a single image. Oftentimes, human observers interpret static scenes with ease, attesting to the fact that under a broad range of circumstances, a single image does provide enough information for semantic-based analysis and understanding. To maintain a sharp focus, this paper is primarily concerned with natural scenes conveyed in single 2D images and does not consider in any detail higher dimensional data such as video sequences (for a recent review, see [154, 168, 86, 169]).

1.2 Focus of Discussion

The main focus of this paper is drawn on computational approaches which seek decomposition of a 2D image in terms of visual components that are intuitively meaningful to humans. In particular, these approaches are concerned with image segmentation based on natural image semantics that captures the general characteristics of a broad range of perceptual objects subsumed under the same category. At the core of current developments of natural image segmentation is

[1]This notion of natural images is not unique to this paper, but is often used, though sometimes implicitly, in the literature of semantic-based image analysis (for an example, see [82]).

a set of conceptual and computational issues concerning the relations between visual semantics and scene segregation. It is a common view that representing the semantics of natural scenes as well as the assumptions/heuristics concerning the nature and structure of visual input is a problem of image modeling; and finding a set of subimages that provides the most meaningful decomposition of a scene according to a particular model is a problem of inference.

Image data can be mapped to these categories either deterministically or probabilistically. The former seeks a model of the underlying processes and structures inherent in scene/image composition to ensure certainty and to remove ambiguity in image partition, whereas the latter consider the composition of visual input as stochastic processes. The diversity and ambiguity in visual appearance of perceptual categories are captured as part of the stochastic causes that are likely to have generated the observation.

The recent surge in interest in probabilistic approaches to segmenting an image into object classes constitutes the main focus of this paper, which seeks to throw light upon the semantic nature of the problem. In this paper, these approaches are collectively referred to as probabilistic approaches to natural image segmentation, or probabilistic approaches in short. A decade of exploration has produced a body of literature of its own, and calls for a thorough and thoughtful review of their theoretical and computational underpinnings. Unlike the other approaches, however, there is a lack of a systematic account of these approaches as a whole from the perspective of scene segregation and their implications for further directions of image segmentation research.

1.3 Organization of Discussion

The account given in this paper starts with a set of conceptual underpinnings of image segmentation. In particular, it highlights the nature and the challenges of natural image segmentation to the conventional frameworks, which seek to delineate significant subimages by visually-defined, content-independent structures – these approaches are referred to in this paper as image-based segmentation. It also covers the shift of attention of a number of approaches to the semantic aspects of the problem. The focus of discussion then shifts to defining the problem of category-based segmentation with a general account of the theoretical formulation of the problem in the probabilistic approaches.

Section 2.3 defines the problems and introduces the major ideas and conceptual principles of probabilistic approaches. It also emphasizes the major distinctions among different approaches in how natural scene segregation problems can be represented under probabilistic frameworks.

The major focus of Section 4 is drawn to the problems of semantic modeling in object class segmentation. It emphasizes the conceptual issues pertaining to the problems, viz, how to capture semantics from visual features, how to integrate image-based and semantic-based visual cues, and how to organize relevant information into a coherent system of representation and inference.

Finally, the last section considers a number of conceptual and empirical issues of the prevailing models in the current developments, many of which remain open and should be addressed in future research into natural scene segregation.

The appendix provides a summary of some major probabilistic approaches to

natural scene segmentation.

2 Image Segmentation: Conceptual Principles

Image segmentation refers to the formation of subimages that decompose a scene. Different classes of subimage formation are found in the literature. Some approaches seek image decomposition corresponding uniquely to semantically defined categories, usually physical objects, embedded in the visual input, whereas the others define regions that may not align with the embedding of any semantic category or physical objects in the image [151]. The latter perspective, which gives rise to image-based segmentation, is the main focus of earlier research in image segmentation and remains a problem of vital importance that spawns much active research. The changing demands of visual applications pose great challenges to image-based image segmentation, motivating a shift in perspective toward semantics-based segmentation, the problem with which this paper is concerned. Segmentation, which organizes visual input into a coherent interpretation, relies on the concepts of coherence and visual organization. It is these fundamental concepts that differentiate classes of segmentation approaches. A comprehensive review of this development is beyond the scope of the paper. Instead, a few representative perspectives[2] will be mentioned in order to illustrate the general persuasion of the image-based perspective on the problem and the reasons for the shift to more

[2]There are a number of good reviews on various aspects of image-based segmentation, such as [136, 21, 118, 142, 37, 38, 150, 27, 32, 52, 151]. For compactness of exposition, no separate citations of these reviews are made in the ensuing discussion.

semantics-based standpoints. These accounts emphasize the basic intuition rather than the formal details.

2.1 Image-Based Segmentation

Image segmentation has conventionally been described as exhaustive partitioning of visual input into subimages over spatially contiguous regions, each of which is characterized by some measure of visual homogeneity and a significance contrast with its surround relative to the interior variation. The partitioning can be thought of as representing image-based visual structure [1]. This view of segmentation has given rise to a manifold of segmentation techniques, which are traditionally discussed in terms of the paradigmatic distinctions: edge/boundary-based, region-based and clustering-based segmentation [40, 138, 62]. Edge/boundary-based segmentation techniques locate the interface between subimages along the visual discontinuities where abrupt changes in image features occur. In contrast, region-based segmentation seeks spatially contiguous homogeneous subimages according to some measure of image properties. Cluster-based techniques group visual input in some characteristic feature spaces into clusters. Subimages are thus defined in terms of two major criteria: spatial contiguity and visual similarity (or dissimilarity). Cross-fertilization between paradigms has also been explored to integrate different tools and to exploit the complementarity between region-based and boundary-based information [38].

A number of mathematical frameworks have attracted a great deal of attention

and have generated very active research to date. Closely related to semantic-based segmentation in general, and the probabilistic approaches in particular, are the well-studied paradigms that formulate image segmentation explicitly as a problem of optimization. Underlying the optimization problem is the concept of energy-minimizing segmentation. An energy function is defined over all possible ways of segmentation, that embeds a cost measure according to some desired properties of good segmentation [141]. The task of image segmentation is defined as the search for the optimal solution that corresponds to the minimum of the energy function. This basic idea resonates across different paradigms, probabilistic and deterministic, throughout this paper. Energy based models are distinguished by the type of energy function and the corresponding optimization technique applied to the models. In particular, the underlying representation gives rise to the class of spatial continuous and spatial discrete models [19, 31]. The following is a brief account of the key ideas of these approaches to highlight the common working principles of different segmentation paradigms despite their distinct standpoints.

The segmentation problem may be formulated in the domain of continuous functions and solved by variational techniques and gradient descent dynamics that evolve the contours in the direction of negative energy gradient according to a set of partial differential equations (PDE's). Two major classes of deformable models – the parametric deformable models, such as snakes, and the geometric deformable models based on the level set function – have been developed to adapt an initially given contour dynamically to the image input

until it stops at the boundary of some optimal subimages by optimizing some energy functions [67, 29, 63, 26, 96, 27]; More generalized formulations have been explored to capture the synergy of gradient-based boundary costs and interior or regional costs in driving the segmentation processes. The Mumford-Shah image energy model (sometimes referred as Mumford-Shah functional) approximates an image as consisting of disjoint homogeneous subimages and piecewise continuous boundaries. Many formulations of the ideas of piecewise smooth segmentation have been independently developed or subsequently elaborated, such as the weak membrane/plate models, the region competition method and many others [12, 116, 120, 171, 26, 27]. Many of these approaches open up the possibility of introducing into variational approaches the statistical properties of semantics-motivated objects in the feature spaces of color, texture and shape; more is discussed in due course.

Discrete approaches give rise to a graph-based formulation of the problem which seek optimal solutions corresponding to the lowest energy partition of a graph. Those approaches seek the configuration that corresponds to the minimum of an energy function, which is defined over an undirected weighted graph to encode observation and segmentation criteria. Many energy functions used in the literature can be summarized in the standard form [161, 73, 72]: $\Psi = \Psi_{image} + \Psi_{prior}$, where the image-driven component, Ψ_{image}, is a function derived from observed data; in particular, when viewed from the image labeling perspective, it measures the cost of a particular assignment given the state or observation. The prior energy, Ψ_{prior}, encodes the segmentation criteria and

the corresponding interpretive constraints used for segmentation. Energy functions may also encode a probabilistic measure or distribution over the space of possible configurations (or assignments) to induce a stochastic random field. The energy surfaces in most real-life image modeling, probabilistic or otherwise, are non-convex with potentially very complex topology. The search for the optima of these surfaces is, in general, extremely challenging. One classical approach is to explore the solution space for an estimate of the optima with Markov chain dynamics [41, 42, 158, 159, 4]. The data driven Markov chain Monte Carlo (DDMCMC) method, for instance, uses image-based visual cues, such as edge or color distributions, to guide configuration formation by means of revision of boundaries, splitting and merging of regions and switching of region models [158, 159]. For other stochastic search approaches applied to partitioning the weighted graph, see [41, 42, 4]. Min-cut/max-flow optimization techniques constitute a widely-adopted alternative to stochastic search approaches in partitioning a weighted graph [46, 166, 20]. A cut is a partition of the sites (nodes) of a graph into two disjoint sets; its cost is the sum of the cost of all edges crossing the cut. The key idea of the min-cut method is to partition a weighted graph with a cut at the minimum cost. To generate clusters with minimum similarity between groups and maximum similarity within each group, the normalized cut approach measures the cost of a cut in terms of both the sum of edge weights across the cut between two regions and the total connection between the sites in either of them with the rest of the graph [146, 148, 147]. The application of these frameworks to semantic-based

segmentation is discussed later.

2.2 The Semantic Gap and Scene Segregation

Image-based segmentation seeks to segregate subimages according to some domain-independent and visually defined concepts of visual coherence. These image-based criteria capture important local constraints, and therefore provide an otherwise ill-defined problem with some minimum structure so that the relevant classes of coherent visual patterns can be recovered by algorithmic means. The simplest yet still powerful of these properties includes the assumption of piecewise homogeneous, smooth or continuous surfaces. More elaborate criteria that incorporate statistical properties in complex feature spaces have been developed yielding significant improvement in the quality of segmentation. It has also been realized that this segmentation paradigm has reached a performance level, where many state-of-the-art approaches tend to converge [157, 43]. Yet, object segregation remains a challenge for image-based segmentation which provides no immediate solution to the problem. The discrepancy between the results of state-of-the-art (image-based) segmentation approaches and human perception is striking, as illustrated in Figure 1. In general, an image is more than a combination of image-based feature patches. Rather, humans perceive meaningful scene composition and they segregate the subimages accordingly. Oftentimes, segmentation boundaries alone provide adequate information for interpreting a scene in terms of semantic categories with which humans usually employ to think of their environment. Despite individual

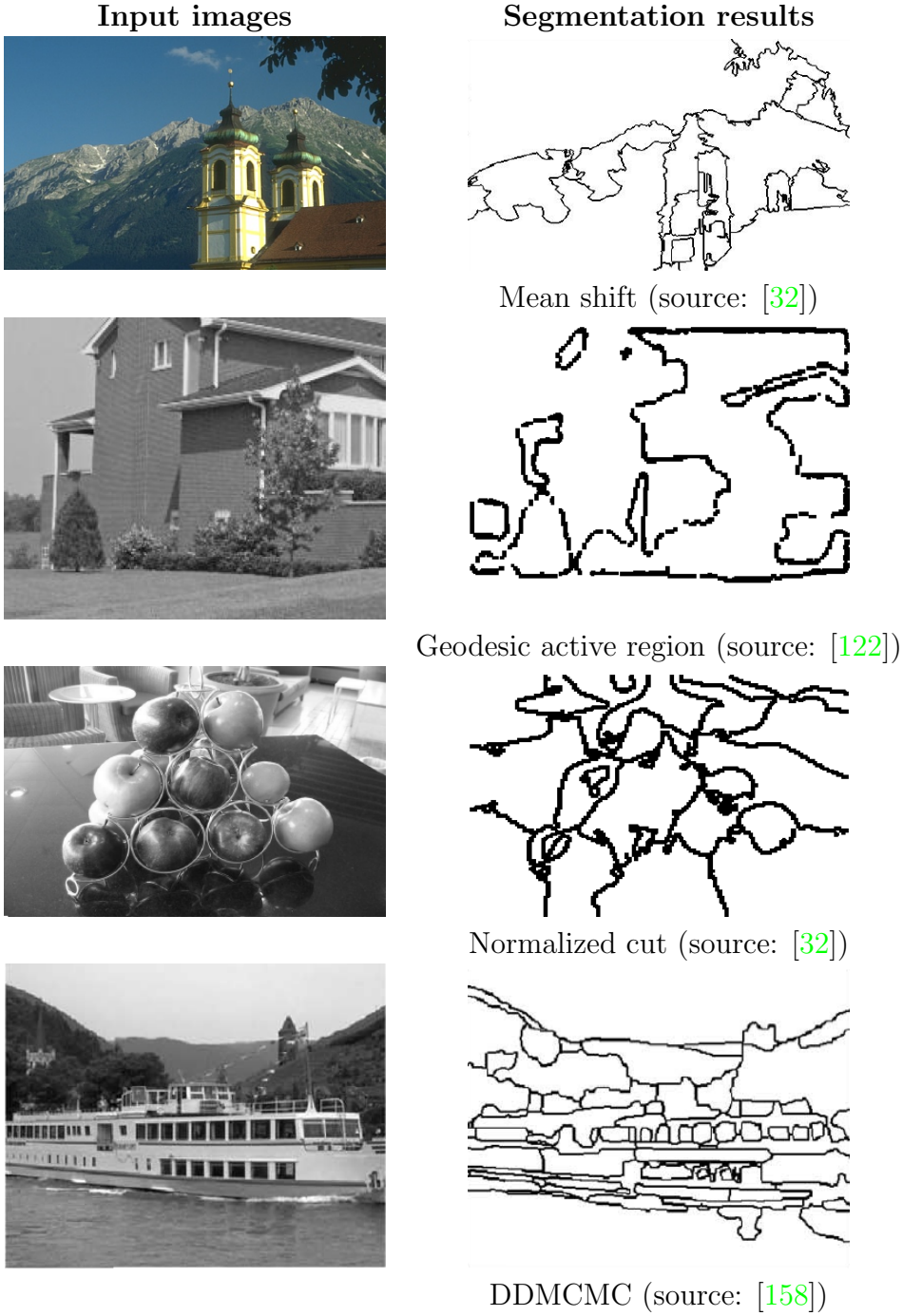


Figure 1: Discrepancies between image-driven segmentation and human perception: some illustrations. From top to bottom respectively : mean shift, normalized cut, data driven MCMC and coupled geodesic active region.

differences, manual segmentation by human observers, as illustrated by the examples in Figure 2, segregates conceptually motivated, semantic-based objects according to their perception of the world.

Strong correlation between the semantic-based description of image contents and the coherent properties captured by the segmentation criteria is required for any segmentation technique that is intended to generate a partition that is meaningful to human perception. As the semantic distinction pertinent to visual perception is extraneous to image-based segmentation criteria, their correlation is but a pleasant coincidence. The discrepancies between image-based segmentation and the semantic contents of images find their causes in the fundamental challenge of the semantic gap.

Research into content-based image analysis demonstrates that semantic concepts which define the categories of our perceptual world, stand in no direct relationship to image-based attributes [110, 167, 45, 97]. Indeed, the issues of lack of coincidence between visual features extractable from image data and semantic-based interpretation of a visual scene – that is the semantic gap – has been raised since the earlier history of semantic-based image understanding and segmentation [152, 65]. In hindsight, it should come as no surprise. Otherwise, ideas, abstraction, and experience would play no role in visual understanding; were this true, given input alone would dictate how humans interpret their visual world, and visual understanding would stand apart from the general faculty of human understanding. The absence of strong relations between two sides of the gap does not imply a lack of visual regularities

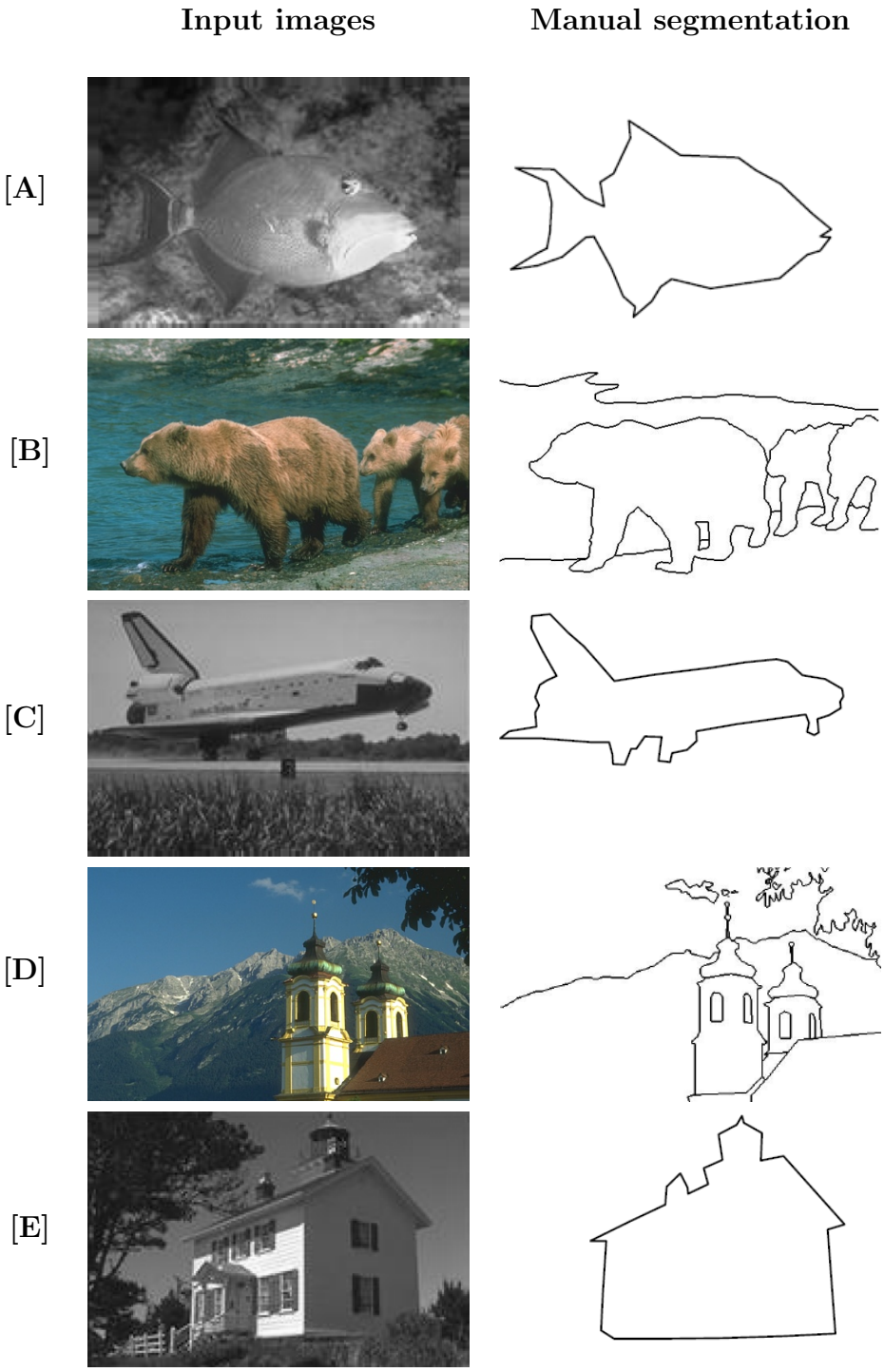


Figure 2: Object segregation based on human perception. Sources for [A],[C] and [E]: [43] and for [B] and [D]: [32].

among members of semantic categories; otherwise, semantic-based visual understanding would be impossible. Rather, these categories are characterized by their variability in visual appearance. Moreover, more than often, their visual appearance cannot be directly explained by the very definition of the categories. A simple object class, such as human faces and horses, appears enormously different across individual instances and across the full range of viewing conditions – not to mention buildings, fruits, animals, flame/fire, rainy days, tools and other more abstract categories. Furthermore, objects defined in terms of the same semantic meaning do not necessarily cluster in clearly delineated groups vis-à-vis other objects in the image-based feature space. The challenge to semantic-based image analysis and understanding is very much compounded by the complexity of scene composition due to occlusion, cluttering, reflection, shading and other unfavorable conditions. The recent surge in interest in semantic-based image analysis and understanding refocuses much attention to the problem of capturing the semantic meaning of images, and motivates active research to discover perceptually based correlations between the semantic-based and image-based descriptors [110, 28, 121].

2.3 Beyond Image-Based Segmentation

As highlighted in the preceding discussion, image-based segmentation emphasizes the similarity of image-based features but considers no semantic distinctions of perceptually coherent parts. This limitation of image-based segmentation as well as image understanding has long been pointed out by many

researchers [40, 61, 2, 155, 9]. In spite of different theoretical emphases, object segregation is always an important inspiration for image segmentation. Indeed, many visual tasks do require segmentation information related to semantic object segregation. Different themes can be discerned in the literature to incorporate perceptual properties that are not directly recovered by measures of visual homogeneity and spatial proximity. One of them, the probabilistic approach, is the focus of the subsequent sections of paper. The rest of this section outlines some other major approaches to perceptual/semantic object segmentation.

Many approaches take the obvious transition to narrow the gap by incorporating object-related properties into the existing image-based segmentation models – including but not restricted to the thresholding segmentation [151], deformable models [35, 63, 122, 30, 123, 27, 70, 22, 31] and graph-based models [47, 48, 49, 50, 51]. Instead of some simple measure of similarity in feature spaces, regions are clustered and visual input is classified according to the statistical properties of visual appearance of the target objects. Much research has advanced along this direction in domain-specific image analysis – a noticeable line of development being segmentation of medical or biological objects.

Many interactive tools have been developed to allow on-the-fly human input that provides information about the target objects for initialization and subsequent modification of the segmentation. Boundary-based methods, such as the intelligent scissors [112, 113, 115, 111], lazy snapping [95], JetStream [124], and others, trace out curves by integrating image-based segmentation criteria and

on-line user-defined information to enclose a target object. Region based approaches, in different segmentation formulations, such as region-growing [114], Markov random field [11] and graph cut [139], extract target objects according to the region statistics recovered from a set of exemplar regions specified by users as foreground or background.

Common to many approaches that attempt more automatic segmentation of natural scenes based on semantic information regarding target objects is the strategy of employing pattern recognition techniques to capture the visual distinctions between object-embedding subimages. Neural networks, Bayesian classifiers, fuzzy models, k-nearest neighbor algorithms and other techniques are frequently used for classifying visual input in terms of visual patterns pertinent to domain-specific categories; for a review of these approaches used in medical images analysis and remote sensing, see [8] and [98] respectively. Similar approaches are applied to natural image retrieval. Blobworld [25, 24] extracts coherent regions according to a mixture model defined in a complex feature space. A number of object-specific segmentation approaches has also been developed.

3 Probabilistic Pattern Approaches to Object Class Segmentation

A probabilistic model is a mathematical description of an inference problem. As a more elaborated discussion is provided in [93], it suffices to state the cen-

tral idea of probabilistic approaches: the problem of object class segmentation is to infer a plausible description, $\mathcal{A}$, in terms of a set of object class labels, with respect to the probabilistic model or probability distribution[3], $pr(\mathcal{A}|\mathcal{D})$, given observations, $\mathcal{D}$. The underlying model encodes the "true value[4]" of each interpretation, i.e., how well each interpretation explains the observation. In probabilistic approaches, these models defined over a set of parameters capture belief about plausible configurations, encoding prior knowledge and all information derived from past observation.

Inference can therefore be viewed in terms of two problems: interpretation selection and model selection. The primary interest of interpretation selection in the context of object class segmentation is an optimal partition that incurs the least expected risk due to misclassification, according to a trained probabilistic model. Optimal decisions depend on the selection criteria. It is common in the literature to seek the mode of the posterior distribution, also called maximum *a posteriori* (MAP), i.e.,

$$\hat{\mathcal{A}}^{MAP} = \arg\max_{\mathcal{A}} pr(\mathcal{A}|\mathcal{D}) \tag{1}$$

or the maximum posterior marginals (MPM's), i.e., for each labeling site $\mathcal{S}_i$,

$$\hat{\mathcal{A}}_i^{MPM} = \arg\max_{\mathcal{A}_i} pr(\mathcal{A}_i|\mathcal{D}), \tag{2}$$

[3]In this paper, probability distribution refers to probability density function of continuous variables and probability mass function for the discrete cases.

[4]This is an interpretation adopted from Hinton and Sejnowski [56].

where $\mathcal{A}_i$ is the description of the subimage i.

The trained probabilistic model is given in model selection which seeks a configuration of model parameters that best adapt a probability model to a set of training data. It is common to compute the optimal configurations of hidden variables and model parameters either under some iterative framework [80, 79, 164, 55] or with some stochastic search procedure [134, 157, 156, 159] using Monte Carlo methods. The idea of object class segmentation as a problem of visual inference plays a key role in probabilistic approaches and motivates much active research on the technique of visual inference for image labeling. Further discussion is presented in [93].

3.1 Major Approaches

To capture coherent and meaningful interpretations of the observed signal, probabilistic models encode the prior knowledge of model semantics and empirical evidence distributed across space over different levels of abstraction. Different approaches can be distinguished in the literature as to how these probabilistic descriptions should be specified. The inference problem of object class segmentation may be solved by searching for the most coherent partition that agrees with the observation. This basic principle gives rise to the strategy of partitioning via stochastic search. Instead of an explicit definition, an image partition may be implicitly given through predicting the object class membership of individual visual inputs. In practice, object class labeling is inferred by means of classification through analysis of local measurement. For the purpose

of exposition, the literature may be organized according to different classification schemes. For clarity of exposition, this paper summarizes the literature using the presentational scheme as illustrated in Figure 3. This scheme should be understood as only a road map for discussion of the literature.

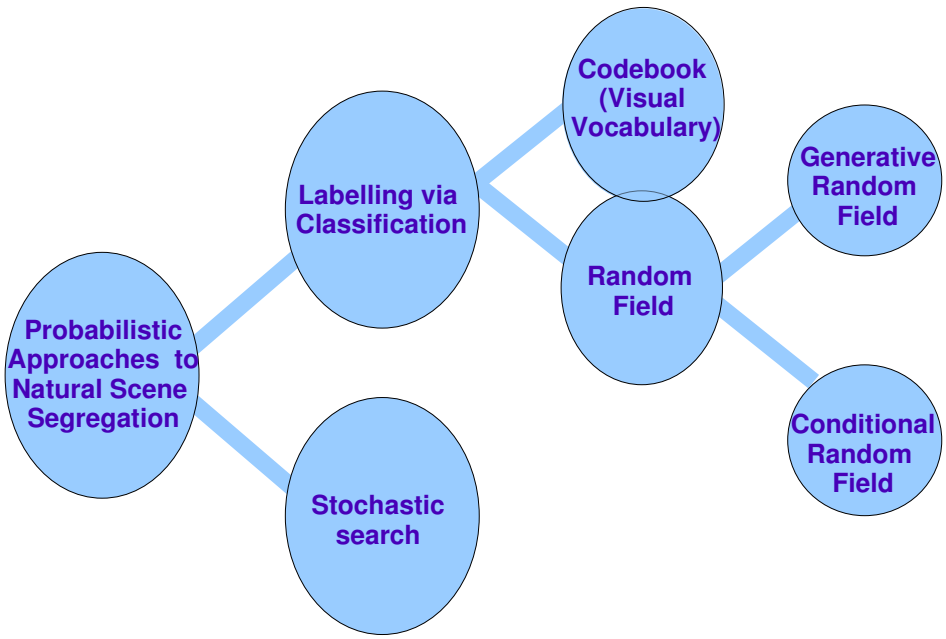


Figure 3: Organization of discussion of different probabilistic approaches to scene segmentation in this paper. The probabilistic pattern approaches are grouped into subclasses in terms of modeling techniques from left to right. See text for details.

3.1.1 Image Partitioning via Stochastic Search The basic idea of stochastic search is to look for the most coherent interpretation according to the posterior model. This perspective adopts the classical definition of image segmentation, which is a configuration of non-overlapping regions: $\{\mathcal{R}_i\}$, that

cover an image, $\mathcal{I}$.

$$\mathcal{I} = \bigcup_{k}^{i=1} \mathcal{R}_i, \quad \mathcal{R}_i \cap \mathcal{R}_j \neq \emptyset, \quad \forall i \neq j. \tag{3}$$

The space of all admissible partitions is a set of all possible configurations of regions that partition an image. The search of the configuration that best explains the observed data may start from an initial configuration and follow a sequence of reversible moves to explore the space under a set of operations for configuration formation/transformation. These operations may include merging or splitting of regions, modification of region description, boundary evolution and others. Illustrated in Figure 4 are three different examples of admissible image partitions.

Figure 4: The space of image partition can be conceptualized as a set of different configurations of non-overlapping regions that cover an image. The configurations i, j, and k, for instance, represent three different elements in the space of admissible partitions. Different elements in the space can be transformed from one to the others under a set of operations of configuration modification such as changing the number, combination, placement, and description of constitutive regions.

The most sophisticated expression of the partitioning via stochastic search strategy is found in the generative framework of image parsing, an extension

of data-driven Markov chain Monte Carlo (DDMCMC) to visual segmentation, grouping, and recognition tasks [158, 156, 157, 159]. An image is parsed into a configuration of constitutive visual patterns with a graph based representation called a "parsing graph." The structure of the graph specifies the spatial

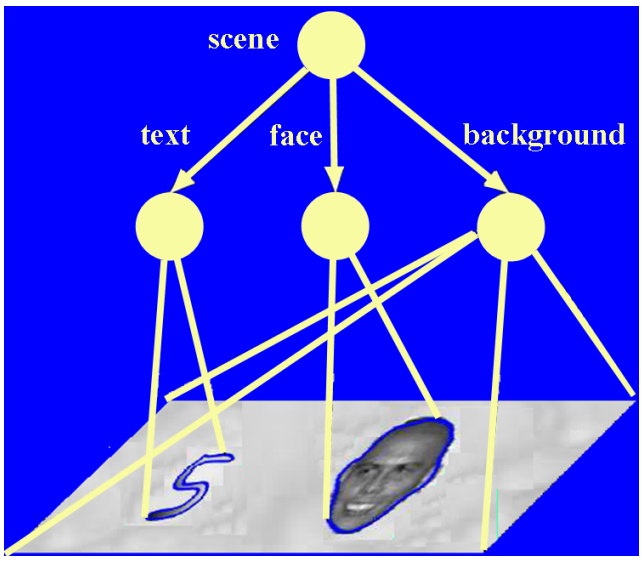


Figure 5: Abstract representation of a parsing graph. The parsing graph takes the form of a three-level tree. The root represents the scene. The intermediate nodes represent the pattern-based regions, which collectively constitute the description of the scene. Each of these region-level nodes is connected to the leaves corresponding to the pixels that form the region.

relationship of the patterns, and the state at each site represents a particular pattern model drawn from a set of pattern families, generic as well as class-specific. The simplest parsing graph takes the form of a three-level tree. As illustrated in Figure 5, the tree consists of a root representing the scene, a set of intermediate sites of visual patterns and the leaf sites corresponding to image pixels. The layer of intermediate sites defines a partition of the image,

where the size of the layer is not known *a priori* and must be estimated.

A parsing tree, $\mathcal{W}$, describes the scene in terms of a set of visual patterns. The objective of the segmentation process is to select the parsing tree such that

$$\hat{\mathcal{W}} = \arg\max_{\mathcal{W}} pr(\mathcal{W}|\mathbf{I}) = \arg\max_{\mathcal{W}} pr(\mathbf{I}|\mathcal{W})pr(\mathcal{W}), \tag{4}$$

where $\mathbf{I}$ denotes observation. Let K^ℓ be the number of visual patterns from family ℓ, and $\{\mathcal{W}_i^\ell\}$ the set of visual patterns of the family. The prior model of the underlying interpretation is given by

$$pr(\mathcal{W}) = \prod_{\ell}^{\rho} pr(K^\ell) \prod_{i}^{K^\ell} pr(\mathcal{W}_i^\ell). \tag{5}$$

Let $\mathcal{V}_i^\ell$ be the region corresponding to the visual pattern $\mathcal{W}_i^\ell$. Visual patterns in different regions are modeled as independent stochastic processes specified by the visual pattern class ℓ. The likelihood of image data is given by

$$pr(\mathbf{I}|\mathcal{W}) = \prod_{\ell}^{\rho} \prod_{j}^{K^\ell} pr(\mathcal{V}_j^\ell|\mathcal{W}_j^\ell). \tag{6}$$

The exact form of the prior and the image likelihood depends on the choice of visual patterns[5]

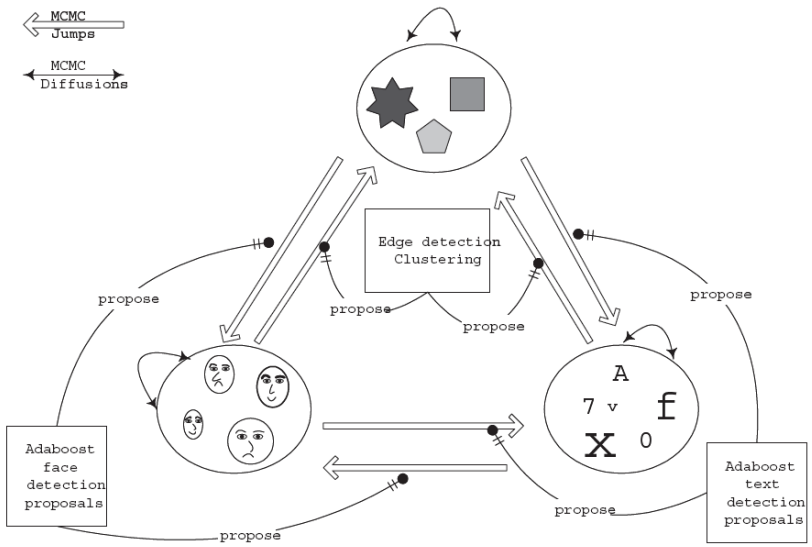
The framework of reversible jump Markov chain Monte Carlo is adopted for searching for the image partition that is most compatible to observation by

[5]Non-class specific pattern families include homogeneous patterns, texture patterns, shading patterns, clutter patterns, and curve patterns while the class-specific pattern families consist of frontal face patterns and text fragments [157, 159].

reconfiguring the parsing graph. The reconfiguration dynamics is governed by a set of graph transformation operators, which either change the structure of the group or node attributes, including a birth or a death of a visual pattern, splitting and merging of a region, pattern model switching, and boundary evolution. The ergodic and reversible Markov chain in the space of parsing graphs ensures that fair samples are generated from the invariant probability corresponding to the posterior model $pr(\mathcal{W}|\mathbf{I})$. Object class specific detectors are deployed to approximate the probability of graph components conditional on the observation and propose a move in the Markov chain dynamics, as illustrated in Figure 6.

This is one of the most explicit formulations that integrates both generative and discriminative methods to solve the problems of object class segmentation. In this framework, generative models, by virtue of their full account of the data generation, ensure the consistency of interpretation while discriminative methods provide a fast but not necessarily consistent solution to restrict an otherwise very extensive search in the most probable regions.

Image parsing is a theoretically well grounded framework for integrating multiple visual patterns visible from different levels of analysis. In practice, however, only visual patterns with comparatively little variability such as frontal face and text fragments are tested in the reported evaluation, despite the application of discriminative methods to drive the search dynamics. Part of the reasons may lie in the complexity of the generative model and the search procedure that would be required for unconstrained scenes. This partitioning



Source: [157]

Figure 6: Abstract representation of the Markov chain dynamics used in image parsing. Data driven processes based on pattern classification are used to propose a new move in the space of parsing graphs in search for the optimal configuration that explains the data.

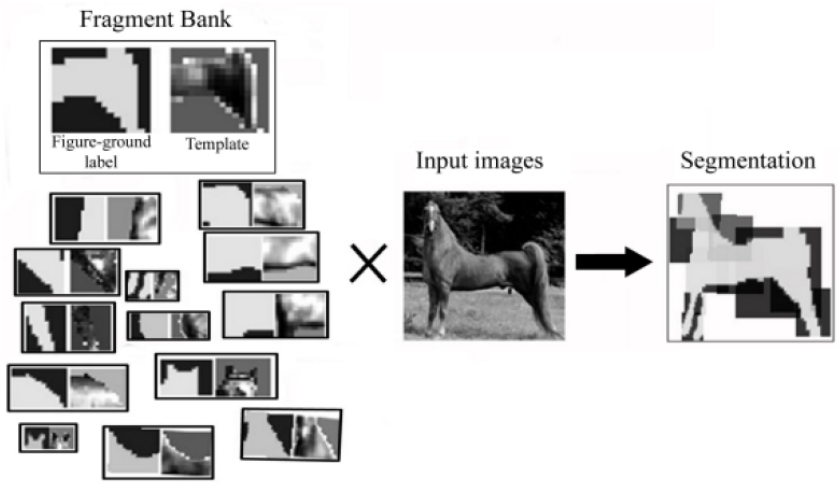
via search principle has also been adopted with a more discriminative flavor to incorporate both image-based and gestalt cues in the decision of partition reconfiguration with moves proposed according to a set of classifier functions [134].

3.1.2 Scene Labeling via Classification A majority of the approaches fall into this category. Instead of directly manipulating the global partitions, these approaches only implicitly define this global solution through predicting the membership of local regions. A widely-adopted choice of representation is the modeling language of conditional random fields which can be interpreted

as networks of classifiers interactively dependent on each other [66]. Other non-graph-based formulations are also possible to implement this strategy. For example, the stochastic properties of the underlying processes may be captured by the empirical distributions observed from the training data or by integrating similarity measures defined in terms of the responses of various classifiers [75, 134].

3.1.2.1 Labeling with visual codebooks The principal idea is to delineate semantic objects by direct recovery of their occurrence in an image from local information. Many approaches define semantic categories in terms of a dictionary or codebook of visual templates. Various computational heuristics have been devised to build a dictionary or codebook by sampling from a training set of interpreted images. The conditional probability of membership given a visual pattern can then be modeled with respect to a similarity metric defined over the dictionary and the membership map associated with each constituent pattern.

The jigsaw approach [15, 14, 17] uses a set of overlapping fragments of an object class with corresponding 'figure-ground' maps to detect class instances in data and to label image pixels. Overcompleteness in representation is the key to this labeling scheme, as illustrated in Figure 7. Given a huge codebook, an object instance in an image is likely to be covered completely by fragments of varying sizes, which are highly overlapped and well distributed across an object instance. Each pixel of the selected fragments can be labeled in terms of their membership in figure or ground with a reliability measure defined by



Source: [15]

Figure 7: Abstract representation of the jigsaw approaches. A library of prototypical intensity patches sampled from a training set is matched to visual input for object detection. Segmentation labels can be assigned according to the figure ground masks associated to the intensity patches recovered from the data.

the hit rate of fragments.

To capture the common characteristics within the class, the fragment library (or dictionary) of intensity patches should be strongly correlated with object-embedding subimages, and low in similarity measures with other subimages, which contains no class instance. According to the Neyman-Pearson criteria, or mutual information criteria, a subset of k most informative fragments is selected from a group of varying sized image patches randomly sampled from a training set that consists of both class and non-class images.

Implicit shape models (ISM) [89, 87, 92, 90, 91, 88] combines the codebook

approach with spatial configuration modeling. The object class-specific model is generated from a training set with segmentation information, based on a discriminative set of prototypical patterns of class instances. Visually similar features are grouped according to a similarity measure and the centers of the resulting clusters are included in the codebook. In particular, the local descriptions are grouped using agglomerative clustering. The similarity between two patches, α and β, is measured by Normalized Greyscale Correlation:

$$NGC(\alpha,\beta) = \frac{\sum\limits_{\substack{\alpha_i\in\alpha \\ \beta_i\in\beta}} (\alpha_i - \bar{\alpha_i})(\beta_i - \bar{\beta_i})}{\sqrt{\sum\limits_i (\alpha_i - \bar{\alpha_i})^2 \sum\limits_i (\beta_i - \bar{\beta_i})^2}}, \tag{7}$$

and the similarity measure between two clusters c_i and c_j, two clusters of local descriptions, is defined as the follows:

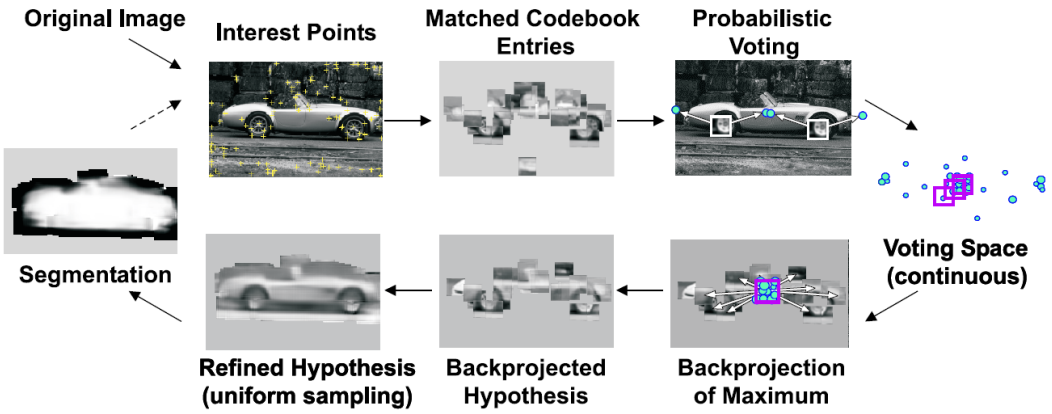
$$\mathfrak{S}(c_i, c_j) = \frac{\sum\limits_{\substack{\hat{c}\in c_i \\ \tilde{c}\in c_j}} NGC(\hat{c}, \tilde{c})}{|c_i||c_j|}. \tag{8}$$

Clusters are recursively merged whenever the similarity between their constitutive members is above a threshold t.

Corresponding to each of these codebook templates is a learned probability of object locations and figure-ground labeling masks. Thus, the codebook provides an object-centered representation of figure-ground assignment. The probability of spatial configurations is approximated by the empirical distribution by searching over the training set for all of the similar occurrences of

the codebook entries, and the positions of the activated entries are stored with respect to an object center. This center-adjusted location can be sampled by a kernel density estimator to obtain a non-parametric probability density estimation of the spatial distribution of codebook entries. A segmentation mask is generated for every occurrence position of each codebook entry.

Object class segmentation is framed as a problem of figure-ground labeling according to a detection-labeling strategy, as illustrated in Figure 8. In the



Source: [92]

Figure 8: Segmentation based on an implicit shape model is characterized by a detection-labeling strategy. Image patches are sampled from an image and compared to the codebook. Matching patches then cast probabilistic votes, which lead to object hypotheses. Segmentation labels can be determined based on the figure-ground masks of those recovered patches that are consistent to the object hypothesis.

object detection phase, intensity patches are sampled from the test image and measured in terms of its similarity with the codebook templates. The activated templates then cast vote for the possible positions of an object occurrence. The

occurrence of object instances, defined by the object class-object center pair, (c, x), can be determined at the local maxima in the voting space using mean-shift model estimation. This voting strategy can be interpreted as a Parzen window probability density estimation for the correct object location.

Since a binary mask of figure-ground label is associated with each codebook entry at a given occurrence location, x, the figure-ground label can be assigned to each pixel $\mathbf{P}$ in the activated regions according to the likelihood ratio of the figure probability to the ground probability:

$$L = \frac{pr(\mathbf{p} = figure|c, x)}{pr(\mathbf{p} = ground|c, x)}. \tag{9}$$

Both the figure and ground probability can be estimated from the segmentation masks associated to the relevant set of codebook entries. The figure probability also provide a measure of confidence in the segmentation results. The segmentation results can be used in a further hypothesis selection stage to reduce the false positives that may arise in the voting process. The best combination of hypothesis, that minimizes the total description length for image, model and error, is selected.

ISM captures the spatial structure of visual patterns of an object class, providing a flexible representation for class-specified object detection and segmentation. Its object-centered representation allows delineating individual objects even under the effects of occlusion in a cluttering scene. The requirement of the approach for a huge codebook, however, restricts its applicability to com-

plex object classes or complex scenes. The segmentation is not accurate for object parts, such as the roof of a car or the head of a cow, which may contain nontrivial inter-class variation or which may be characterized by complex configurations of fine structures.

The topic (or aspect) recovery approaches[6] also use a codebook of discriminative visual patterns to infer the latent topics of each subimage extracted by image-based techniques [140, 162, 23]. The key idea is to explain observation in a high dimensional and usually sparse feature space by a set of latent topics populated in a lower dimensional probabilistic semantic space. The joint distribution of observation is factorized into a product of local condition distributions of visual patterns given the latent topics and the distribution of the topics. In the context of object class segmentation, each subimage, usually an over-segmented superpixel, is modeled as a mixture of the latent topics. These latent topics are inferred from a set of discriminative visual patterns extracted by image-based techniques [140, 162, 53]. The recovered topics are then mapped to object class membership.

3.1.2.2 Labeling with graph-based representations Graph-based representations are common frameworks among different approaches to object class segmentation. The cornerstone of these approaches is the graphical structure of representation and inference. In general, the probability distribution of scene

[6]Topic recovery approaches, also known as latent class approaches, refer to those inference approaches that extend the probabilistic latent Semantic analysis (PLSA) [58, 59] or latent Dirichlet allocation (LDA) [13] to semantic-based visual analysis. LDA can be viewed as an extension of PLSA with an additional Dirichlet prior for the probability distribution of the latent topics.

labeling in terms of visual categories is defined over a graphical structure of representation to encode both observation and *a priori* visual constraints in terms of spatial configurations of filter/classifier responses. The main purpose of graphical representations is to capture by a probability distribution the spatial structures of image features and computational decisions among different sites. These structures are captured by the decomposition of the joint distribution in terms of the neighborhood structure of the underlying graph, such that, each site depends on only a subset of sites in a local neighborhood. That is, given the state of its neighborhood, the site is statistically independent of the rest of the graph.

The probabilistic approaches in the literature of object class segmentation are increasingly characterized by a common language developed in stochastic field theory. They are characterized by modeling a random field defined over an undirected graph, where the sites (or nodes) represents random variables defined over a space of states. These models encode the statistical regularities in the spatial configuration of local measurement, object class membership assignment as well as intermediate explanatory factors. From the modeling point of view, random fields can be viewed as normalized energy-based models. The most well-known class is the Gibbs random field, where the energy, $\Psi = \sum \Psi_C$, defined for each possible configuration is a linear combination of local energies,Ψ_C, each encoding important contextual constraints over the neighboring sites in the clique, C. Many issues in relation to the deployment of random fields in segmentation modeling are the subject of discussion in Sec-

tion 4. For a more detailed account of probabilistic graphical models in object class segmentation, see [93]. It is important to highlight a key distinction in the random field approaches to natural scene labeling: the generative and the conditional random fields.

The application of random fields to image modeling is traditionally associated with the probabilistic generative framework. In object class segmentation, probabilistic generative models encode the input patterns, output configurations, and in many cases, intervening hidden pattern layers of a pattern system by a joint probability distribution [80, 79, 164, 60]. In the absence of any intervening conceptual layers, an ensemble of visual patterns is defined in terms of both an observation model, or likelihood distribution, $pr(\mathcal{D}|\mathcal{A})$, of observation as well as the prior probability distribution, $pr(\mathcal{A})$, of output configurations. The observation model predicts observed states in terms of their generative processes. According to Bayes' law, the posterior model is given by

$$pr(\mathcal{A}|\mathcal{D}) = \frac{pr(\mathcal{D}|\mathcal{A})pr(\mathcal{A})}{pr(\mathcal{D})}, \tag{10}$$

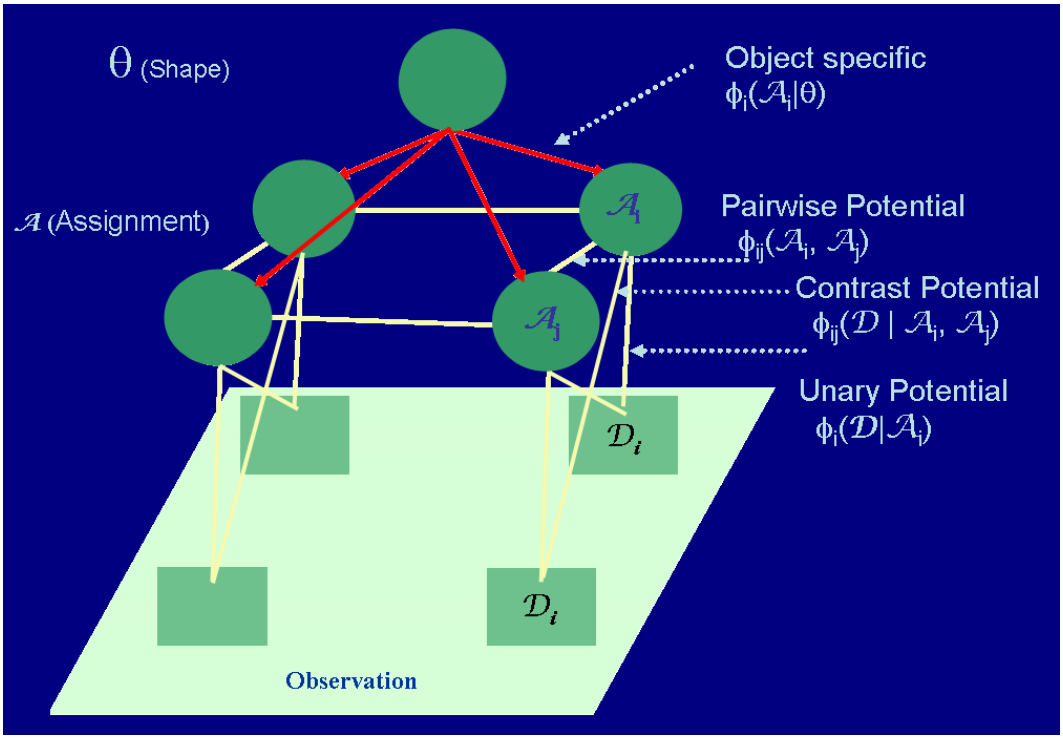
that describes how likely a particular configuration given observation.

Consider as an example the Obj-cut segmentation model [78, 81, 80, 79], a probabilistic generative framework for object class-based foreground extraction. It seeks a configuration of labels drawn from a binary image label set $\mathcal{L} = \{ground, figure\}$, that is most likely under a given probability distribution, $pr(\mathcal{M}|\mathcal{D})$, of configuration given image data. The posterior distribution

of the configuration is given by an object-specific Markov random field (MRF) characterized by an energy function, Ψ, that specifies both image-based and semantic-based constraints over local configurations, as illustrated in Figure 9. A unitary energy component captures the RGB distributions for foreground and background. A pairwise component give preference to same labeling for sites with pixel of similar color. A contrast-sensitive component encourages boundaries of regions to be consistent with image edges. An object specific component assigns low energy to object label for pixels that fall inside the object, given the model parameters of the object class. Given a set of s samples of model parameters $\{\Theta_i\}_{i=1}^{s}$, learned from the training set, a sample-based solution is given by minimizing the following objective function:

$$\widehat{\mathcal{M}} = \arg\min_{\{\mathcal{M}\}} \sum_{i=1}^{s} \Psi(\mathcal{M}, \Theta_i) pr(\Theta_i | \mathcal{X}_o, \mathcal{M}^{\mathbf{T}}). \tag{11}$$

$\widehat{\mathcal{M}}$ can be optimized using MINCUT energy minimization procedure. The objective function given in Eq. (11) is a mixture of experts. Each of these experts contributes its individual opinion according to a particular hypothetical model (or view) of the object class to the combined decision with a weight proportional to the probability of its occurrence defined by the sample parameter set, Θ_i, given the image data and the true configuration. The performance of a 'panel' of experts is found superior to the assignment based on the MAP estimate of the model [80, 79]. Due to the lack of knowledge about the true configuration, this term is approximated by the probability of the parameter

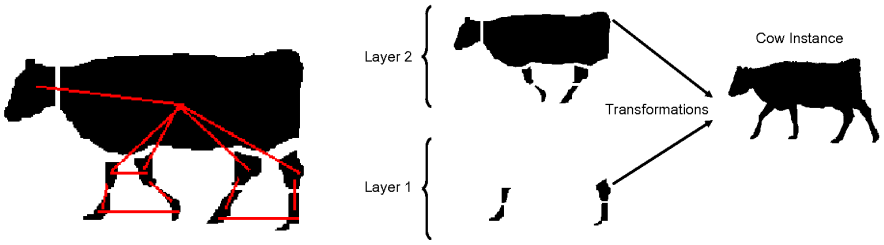


Modified from [80]

Figure 9: Abstract representation of the Obj-cut segmentation model. Obj-cut models figure-ground segmentation by a random field based on four energy components, which encodes color compatibility of a pixel with its label, accounts for the pixel location with respect to the recovered object boundaries, encourages similar labels assigned in local neighborhoods, and ensures figure-ground boundary assigned in the areas of high image contrast. Used for object detection to improve the boundary alignment of segmentation is a layered pictorial structures model, which encodes a shape-based part model of the target object class.

set given image features according to the object class model.

Object class information is provided by a layered pictorial structure (LPS), a MRF with each site corresponding to a 2D pattern that represents a part of



Constellation Object instance generation

Figure 10: Object representation using LSP in Obj Cut

the object class. The random field specifies a generative model for possible shape, appearance and spatial layout of a object class in terms of its rigidly moving components [81]. A part label is defined for a set of pixels that define the part, and the locations, orientations, scales and layer numbers of each pattern in a constellation of parts characterize an object instance.

An object instance is represented as a configuration of extracted parts arranged in layers at relative depths to allow for possible deformation, articulation and self-occlusion, as illustrated in Figure 10. The random field constrains the structure of selected part labels to a set of valid configurations. The LPS is characterized by pairwise constraints between sites according to a Potts model, such that all valid configurations, being considered to be equally likely, are assigned an energy level which is lower than the level associated with invalid configurations. The LPS model is matched to the image to obtain the samples $\{\Theta_i\}$, from the distribution of the model parameters given image features, and the posterior of Θ given image features is approximated using loopy belief propagation (LBP). A object class segmentation is given by the assignment

selected according to Eq. (11).

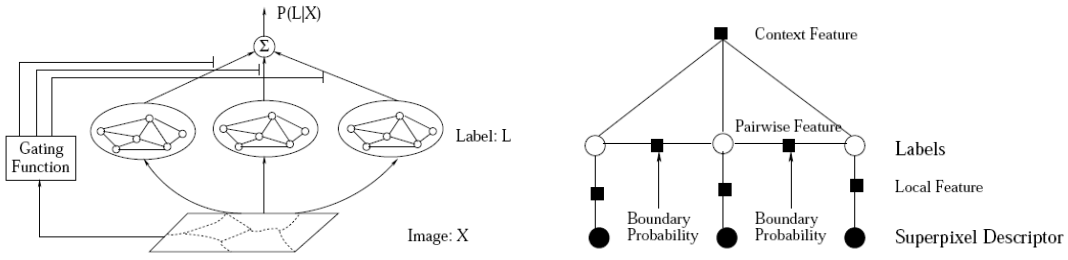
Part parameters, Θ, including number of parts, part masks, location, orientation, scale and appearance of object parts, are learned from a set of training videos, each of which contains an exemplar object instance of the object class in motion [78]. An initial estimation obtained from a set of fragments extracted by clustering rigidly moving points is iteratively refined by optimizing one parameter at a time while keeping the others unchanged until no further reduction of model energy is possible.

The recent proliferation of conditional random fields in object class segmentation highlights the paradigmatic shift towards a more direct (discriminative) approach to the problem, building interpretations around an explicit definition of the conditional probabilistic model, $pr(\mathcal{A}|\mathcal{D})$, instead of a generative model as a whole. Conditional random fields can be interpreted as a network of pattern predictors communicating with each other to exchange information in labeling decision [66]. Many object class specific structures can be captured through a broad range of experts, such as neural and linear classifiers, textons, boundary detectors, shape descriptors, spatial maps, appearance fragments, to name just a few; for examples, see [54, 132, 55, 66, 94, 149, 165, 60, 135].

Instead of briefly mentioning each of these approaches, the following discussion focuses on the scene segmentation framework of the Mixture of Conditional Random Field (MoCRF) [55]. The integration of image-based and semantic-based visual cues is achieved in two steps. First, an image is over-segmented into superpixels[7], each of which is a homogeneous, spatially contiguous region.

A image-based algorithm, such as the normalized cut, can be adapted for partitioning an image into regions of consistent size, which is kept small enough to assume the boundaries of the segments to be consistent with those of the object classes.

The second step is to assign class labels to these superpixels. The image is



Source: [55]

Figure 11: The Mixture of Conditional Random Fields. **A** : A set of context-specific models are applied to the superpixel descriptors for scene segmentation. A gating function is used to modulate the relevance of each context to a given image. **B** : An abstract representation of the model which encodes the rules of segmentation which account for (1) the local information discovered from superpixel descriptors and specific label compatibility; (2) pairwise interactions between labels of neighboring sites, modulated by the boundary probability; and (3) global bias provided by the context-specific average label distribution.

modeled by a mixture of context-dependent conditional random fields (CRF's)

:

$$pr(\mathcal{A}|\mathcal{D}) = \sum_{c \in \mathcal{C}} pr_{_M}(\mathcal{A}|\mathcal{D}, c) pr_{_G}(c|\mathcal{D}) \tag{12}$$

where each $pr_{_M}(\cdot|\cdot, c)$ is a CRF for a context c. The gating function, $pr_{_G}$, generates a probability distribution of scene context, specified by a classifier based on the aggregate statistics of the image date $\mathcal{D}$. Each context-dependent CRF is defined with respect to a graph, $\mathcal{G}$, where each label site corresponds to a superpixel and only those of neighboring superpixels are connected.

Given a given context, a CRF encodes the semantic-based constraints of the label field. In particular, three different kinds of constraints are used in [54]. A classifier is used to measure the probability of an assignment for a single site, given some local features of the underlying superpixel, such as color, texture, and edge information. The interaction between neighboring sites are captured by a pairwise function based on the compatibility of the site labels and a measure of boundary presence between the underlying superpixels. A measure of similarity is incorporated to constrain the overall image label distribution to confirm the relative proportion of the various labels in a typical scene with the given context.

Given the model, an assignment is predicted for a new image according to the Maximum Posterior Marginal (MPM) criterion:

$$\hat{\mathcal{A}}_i = \arg\max_{\mathcal{A}_i} \sum_{c \in \mathcal{C}} pr_{_M}(\mathcal{A}|\mathcal{D}, c) pr_{_G}(c|\mathcal{D}), \tag{13}$$

where the marginal label distributions of each superpixel, $pr_{_M}$, are inferred by loopy belief propagation in every context-dependent CRF's.

[7]Superpixelization is first proposed in [134] as a preprocessing step for learning good segmentation from human segmented data.

Apart from the random field approaches, there are other graph-based frameworks for object class segmentation. The rest of this section provides a brief summary of these formulations of the problem, starting with the directed graphical models. In contrast to the undirected graphical representation of random fields, directed graphical models associated a probability distribution with a directed acyclic graph on a set of random variables, where the joint distribution is given by the product, over all the sites of the graph, of the conditional probability one for each variable conditioned on the variables corresponding to its parents [39, 10]. The common core of topics recovery approaches, mentioned in the preceding discussion, rests on their representation with directed graphical models of the generative processes of observation. Another deployment of a direct graphical model is the belief network which encodes the prior model of scene descriptions across scales of granularity [34]. The description associated with a node in the belief tree is dependent only on the coarser scale description at its parent node given those at all coarser scale nodes.

Graph-based representation plays important roles across different probabilistic pattern approaches to object class segmentation. As mentioned, the stochastic search approaches [158, 134, 156, 157, 159], for instance, seek the most probable segmentation by reconfiguring the underlying graphical models. Similarly, a graph is deployed to represent the conceptual hierarchy of scene decomposition in [103], where an outdoor scene is recursively decomposed into components (each of which corresponds to a node of the graph) with respect to three

types of relations: object composition (is-part-of), object typology (is-kind-of) and spatial relation. A set of rules are extracted from the graph to guide classification.

An object class may also be represented by a tree-based canonical model [153]. Images are represented by trees of subimages, where subimages at the ancestor level correspond to more salient regions than those at the descendant ones. According to the assumption of their frequent occurrence, the object class-specific subtree is extracted from a set of training images under the operations of tree-matching and tree-union. Detection and segmentation of the target object class are achieved by matching the tree of subimage extracted from a test image to the canonical model.

4 Issues of Modeling in Object Class Segmentation

This section provides a brief summary of major strategies proposed in the literature of the probabilistic approaches to address the major issues of modeling in object class segmentation: (1) the measurement problem: how to capture semantics from visual features, (2) the integration problem: how to integrate image-based and semantic-based visual cues, and (3) the organizational problem: how to organize relevant information into a coherent system of representation and inference.

4.1 Measurement Problems

The first problem is concerned with mapping measurements in some feature space to local semantic patterns. The main issues are related to the grounding of semantic representation of image content in visual properties extractable from measurement. Image models based on local intensity patches provide an intuitively straightforward and computationally simple mapping to partition an image in terms of object class membership. This strategy, adopted by many approaches [15, 89, 17, 14, 87, 92, 90, 16, 91, 88, 94, 149] to classify membership of visual inputs, can be effective for certain classes of problems. Among others, two major conditions favor this class of modeling vocabularies. First, the object class subimages can be characterized by a limited set of archetypal intensity patterns, thus not subject to considerable variability due to scene structures, viewing conditions or differences among member instances. Second, the local appearance of different object classes do not share many common features, such that subimages corresponding to different object classes can be easily set apart by distinct distributions over the pattern space. However, these conditions do not hold for natural scenes in general.

A shift away from local appearance fragments to other feature spaces is evident in the literature. The choice of feature space includes but is not restricted to object boundary fragments [134, 107, 80], local intensity/brightness contexts [71, 16, 94, 149, 153, 165, 60, 135], shapes/contours [132, 133, 153, 159, 172], structures [89, 87, 92, 80, 16, 79, 88], colors [103, 34, 54, 164, 55, 94, 18, 23, 135, 163], textures [103, 34, 134, 54, 80, 71, 132, 164, 55, 79, 18, 23, 135, 163]

and edges [54, 132, 157, 164, 55, 94, 149], In spite of the on-going debate about which features are most meaningful for object-categorization and recognition, no privileged feature classes can guarantee the best performance across all object classes under all scene conditions.

Indeed this important and yet difficult problem has been a subject of discussion since the early years of pictorial pattern recognition [21]. In general, texture and color [107] have been found to be most effective, and have been most widely adopted for predicting outdoor scenes dominated by objects in highly variable forms, such as vegetation, buildings, dirt tracks, and the like. On the other hand, geometrical features such as shapes, structures and constellations of parts are generally more reliable cues for objects that are physically generated from some types of blueprints or prototypes such as text, biological forms or artificial objects [92]. These alternatives are to some extent related to the classical choice in image-based segmentation between those employing region-related and those emphasizing edge-related information. In any case, object detection usually involves a complex interpretation of visual stimuli over a context residing in multiple feature spaces. It is common to deploy a combination of expert models, each specialized for a particular set of classification/detection tasks, to incorporate their contributions and to account for their relative merits within an integrated framework of interpretation.

The general strategy of capturing category-specific semantics is to model the conditional probability of semantics-relevant visual patterns given feature measurement. Conditional probability distributions can, in general, be repre-

sented non-parametrically by histograms, which are applied to modeling edges [76, 74, 157, 55, 140, 23], color [75, 55, 135, 23, 163], or texture [75, 134, 80, 157, 79, 23, 163]. The χ^2 distance between two histograms may be interpreted as a probabilistic measure of affinity in local structures between subimages [134, 107]. Similarly, the distances between a subimage histogram and a set of exemplar histograms can be used to capture pattern familiarity [132]. It is also common to reinterpret the confidence measure of a linear combination of classifiers, such as boosting, as the probability distribution of pattern ensembles [157, 149].

Alternatively, the conditional likelihood can be represented by parametrized models. Among the most well-known ones are Gaussian models, (such as for edges [164] and for intensity patterns [157]), and Gaussian mixture models. The latter provide a more versatile way to capture the complexity of the object class-specific pattern ensembles and have been applied to modeling object class patterns in color spaces [164, 149] or texture spaces [164].

Object class-specific visual patterns are characterized by intricate structures of both long-range and short-range dependencies between the responses of visual channels across space. Local descriptions designed to capture statistical regularity over a short range play an important role in object class segmentation. It is not only because of the complexity associated with modeling long range interaction, but also due to the intrinsic character of natural images. Natural scenes are complex, cluttered with objects in various ways. Occlusion gives rise to partial views that may consist of a set of disconnected visual patterns corre-

sponding to the same object. In the absence of any *a priori* knowledge about an object present in an image, visual patterns that characterize the local view of an object class provide the most reliable information for predicting object embedding in the scene. This helps to explain the preference for features that bear no intrinsic relationship to global spatial forms, including color, brightness, local gradients and textures in object categorization [75, 107, 129].

Geometric forms of object classes can be characterized by local descriptors such as contour fragments, orientation energies, local spatial structures or edge groups [99, 107, 80, 132, 133, 79, 159]. For instance, a set of prototypical curves is used in Obj-cut to represent the outline of the parts of an object [81, 78, 80, 79]. Familiar spatial patterns may be modeled by a mixture of Gaussians with a library of prototypical shape pieces, known as 'shapeme' [132, 133, 131]. The global descriptors are eschewed but not absent in natural scene representation [132, 157, 172]. Probabilistic descriptions encoding variability of spatial forms may be helpful for alleviating some degrading effect of occlusion and cluttering. In general, global descriptors of object classes have been found to be ineffective for tasks of semantic-based image interpretation of natural scenes, marred by the degrading effects of occlusion.

There are many approaches [103, 15, 89, 17, 92, 88, 149] that rely solely on the discriminating power of classifiers to provide an object class labeling for a scene. These approaches, on one hand, highlight the predictive power of object class semantic models in object class segmentation, but on the other, turn the problem of segmentation into a by-product of object class-specific subimage

detection and recognition. Yet, it is coming to light that in the absence of low-level image analysis, the performance of object class models declines in effectiveness, accuracy, and reliability as the object classes becomes more complex in form and more variable in appearance [94]. Deployment of *a priori* knowledge of visual semantics is to help removing ambiguities concerning meaningful structure recovery, which cannot be accomplished satisfactorily based solely on image data. These ambiguities are partly due to the highly complex relations between the visual appearance of object class-specific subimages and the semantic definition of the concept. On the other hand, semantic-based models are usually strong in coarse representation that leaves out many details unattended [94].

4.2 Integration Problems

The integration problem arises, which motivates many frameworks to exploit image-based patterns for furnishing a mapping from the semantics of object classes to their pixel level image embedding. An obvious and simple option is to apply the object class model directly to the homogeneous regions obtained from image-based processing [109, 69, 68, 77, 64, 14, 71, 132, 55, 140, 23, 135]. Image-based processes play two roles in forging a representation. First, over-segmentation according to image-based criteria yields 'building pieces', usually referred to as superpixels, each of which is the basic unit for object class membership assignment. A probabilistic model can be defined over some pre-defined partition generated by an existing technique such as graph-cuts

[14, 55, 140, 135, 23], constrained Delaunay triangulation (CDT) [132] or the watershed algorithm [71]. Second, these regions, homogeneous according to some image-based criteria, become the perceptual units to collect statistical measures for object class membership prediction. In the stochastic search framework, superpixel grouping can serve to propose plausible image partitions [134].

Many classification-based models are characterized by conditional random fields over graphs of superpixels. A set of classifiers (multilayer perceptrons) may be deployed to measure the probability of an assignment for a single site, given some local features of the underlying superpixel, such as color, texture, and edge information [55]. The interaction between neighboring sites are captured by a pairwise energy function based on the compatibility of the site labels and a measure of boundary presence between the underlying superpixels. To address the problem of label consistency, a cross-scale framework [14] may be used to represent the pattern-subpattern relations with a tree of superpixels, built by adaptively grouping small collections of pixels into larger ones according to their similarity and saliency [143, 144, 145]. A different probabilistic formulation of this strategy based on a library of binary patches of shape parts can also be found in [16].

By involving image-based segmentation processes in the partition decision, these superpixel-based approaches yield better segmentation. In a sense, the image model is split into image-based and semantic-based representations. Other than region labeling and merging through common labels, the exist-

ing algorithms provide little or no updating mechanisms to revise the initial decision in light of object class interpretation. The notable exception is the integration of codebook-based object class segmentation and the deformable model in an iterative framework [71]. An initial segmentation given by some image-based technique is allowed to be deformed in a sequence of coupling steps of region labeling and region morphing. This approach may be viewed as the deterministic counterpart of stochastic searching approaches.

In the absence of any channel for semantic-based influence on segmentation, the assumption of over-segmentation that all the object boundaries are preserved in the segment boundaries becomes critical and the performance is very sensitive to the quality of the image-based steps. These superpixel-based approaches restrict the contributions of semantic-based interpretation to providing (1) an interpretation to the pre-segmented regions, and (2) semantic-based control over region growing. This can be viewed as an extension of image-based segmentation by introducing object class semantics into its control mechanism that governs region refinement.

Alternative approaches have been taken to encode the statistical properties of the visual scenes or object classes in terms of both image-based and semantic-based visual patterns [54, 80, 79, 157, 18]. A key issue of modeling these pattern ensembles is how to evaluate evidence supplied by different pattern models, both image-based and semantic-based. The Markov field aspect model, for instance, imposes spatial coherence constraints by a random field model on latent topic induction on overlapping image patches [162]. Obj-cut seeks a

segmentation framework that is able to encourage the segmentation to follow the image-based structures, such as edges, but at the same time, to resemble an object [80, 79]. The image processes are encoded with a random field which consists of an object class specific labeling model coupled with a part-detection model to predict figure/ground assignment. The image model is characterized by an energy function that specifies both image-based and semantic-based constraints over local configurations, as discussed in the preceding section. This framework illustrates a common strategy to integrate multiple visual models using a linear combination of energies, each of which encodes a particular class of expected or familiar visual patterns.

4.3 Organizational Problems

It remains a challenge as to how visual patterns should be organized into a perceptually coherent representation that adequately captures the formation of object-embedding subimages. Perpetually meaningful patterns such as junction patterns, contour patterns and curve patterns are helpful for capturing non-local characteristics of object class specific subimages [134, 133]. Image parsing applies the partition strategy to decompose an image into regions and curve structures, such as free curves, and curve groups with (nearly) parallel or tree structures [159]. In a more discriminative fashion, an image partition is generated by grouping superpixels into a good segmentation [134]. Curvilinear continuity of a hypothesized segmentation is defined in terms of change in angles between contour edges at the junction of every pair of adjacent su-

perpixel junctions on the boundary of the segment. In an image labeling framework, visual cues are integrated through a random field characterized by an energy model, which incorporates contextual constraints of grouping regions with strong contour continuity [132].

Hierarchical representations constitute a common framework for recovering visual structures through varying ranges of observation and analysis [34, 14, 54, 135]. Some hierarchical structures can be conveniently captured in a multiscale feature space on the level of object class based classifiers. For instance, regions can be clustered and merged according to an energy representation that matches measurement with an object class-specific template in a wavelet transform domain [77, 64]. Object class-specific descriptions can be expressed in terms of a set of filters applied to different feature spaces in a range of scales by modulating the filter responses with Gaussian kernels of varying scale parameters [75].

Hierarchical architectures are also common, that consist of representations and reasoning in different granularity levels, usually referred to as scales of analysis. Contextual information is captured from different spatial regions, on which the saliency of visual cues that characterize a semantic-based class critically depends. A region that is too small provides insufficient image data, while a region that is too large may contain a mixture of objects and their background.

The multiscale conditional random field (mCRF) method [54] deploys a local classifier to predict an object class-specific label assignment given a set of fea-

tures, measured within an image patch around a pixel. This label field then moves through a hierarchy of models which encode the geometrical relations and patterns of the field over an expanding scope of information integration and analysis. A hierarchical conditional random field is defined in [135] on a forest of multiscale image partitions generated by a graph-based method [33] with varying amount of image smoothing and constraints on the minimum size of regions. The topology of the forest encodes the spatial relations between these segmented regions or superpixels, where a (non-root) node is connected to a single parent with the maximal overlap in the number of pixels between parent and child regions. Other similar approaches including contextual dependent conditional random fields [55] and tree-based fragment segmentation [14] are mentioned previously.

In a subsequent extension of the implicit shape model [92, 90], a codebook of scale varying patches is compiled to capture the appearance variability of object instance over scale. This latter approach represents an attempt to address the problem of scale-invariant object categorization under the probabilistic framework. In general, scale-dependent signatures of object classes have not been taken into account in an explicit and adequate way. Many probabilistic models implicitly assume either a fixed scale for object class modeling or that class instances can be represented adequately by some scale-independent patterns for some given choices of semantic categories. With this aspect of pattern representation largely overlooked, these approaches also forgo the advantages that an invariant representation may bring.

Parsimonious representations introduce unobservable constructs to interpret observation. In object class segmentation, hidden layers of visual patterns may act as representational constructs for encoding *a priori* knowledge such as familiar object class-specific configurations [80, 79], expected structures of occlusion processes [165, 66, 60], geometric/perceptual patterns [159], or scene descriptions [103]. These interpretive structures, though not directly observable, are essential for the detection of object class embeddings in image data. Often, many of these hidden structures capture the underlying invariance of appearance differences due to viewing conditions [80, 164], seasonal/time of day variations [103], intraclass deformation [157] and other effects. Hidden variables also play a central role in disentangling the longer range dependence between observed states, and may be interpreted as intermediate results on the way to the best interpretation [44, 170, 83, 51].

Representations may be organized with layered structures of visual and abstract patterns [80, 164, 79, 159, 165, 66, 60]. The object-specific random field in Obj-cut captures the spatial form of an object class in terms of a composition of 2D patterns, termed parts, organized in layered pictorial structures [80, 79]. Similarly, the located hidden random field assigns object class membership via part labeling [66]. The model energy encodes the part appearance and the local dependence between parts via a set of part classifiers. Layout consistent random field models extend this idea with a set of asymmetric pairwise energy components to encourage local and non-local compatibility between part labels in terms of layout consistency under possible occlusion conditions [165, 60].

An important class of hidden variables captures the image embedding of object class specific patterns in terms of the spatial transformation of the canonical representation of visual appearance. For instance, to encode the putative poses of the object parts, Obj-cut explains the visual patterns of the object class specific pictorial structures in terms of the location, orientation, scale and occlusion of these parts. The configurational constraints of a valid composition of an object instance is encoded by the model energy in terms of the legal ranges allowed for these parameters.

In the located hidden random field model, part-specific patterns are tied to an object-based frame through hidden location variables to encode the plausible spatial configuration [66]. Translation and left/right flips of object instances relative to the canonical representations are encoded to explain the part appearance in layout consistent random field models [165, 60]. In addition to the object-based transformations with respect to translation and scale, LOCUS [164] introduces a hidden deformation field to account for the local variations of image instances from the canonical patterns. The field consists of (1) a set of discrete shifts defined over a set of non-overlapping patches of the image and (2) a prior model that encourages spatial consistency in the field by penalizing the squared difference between changes across neighboring patches. Similarly, image parsing models object class specific patterns with B-spline-based boundary representations varying under the operators of affine transformation and elastic deformation [157].

The hierarchical structure of representations based on prototypical structures

allows the intra-class invariance to be explained by a set of expert modules sensitive to different transformation. Obviously, these strategies for modeling the invariance of visual appearance in object classes has a long history traced back to the common idea shared by these approaches and the deformable template approaches, mentioned previously.

Among the probabilistic approaches to object class segmentation, the deployment of unobservable, explanatory constructs for the most part serves the general strategy of object class detection/labeling based on visual templates and their corresponding filters/classifiers. As a result, deep interpretive hierarchies are rare; they usually involve no more than one or two layers of intermediate structures, closely tied to object class appearance to account for viewpoint changes or other external sources of intra-class variability. The purpose of these detection and classification mechanisms is to recover the image embedding of some prototypical patterns. Yet, the internal structures of pattern organization that give rise to the semantic content of subimages are less explored. This greatly restricts their generalization and predictive power in capturing the inherent semantics of interpretive concepts.

5 Discussion and Conclusion

This section shifts the focus to a more general discussion of the paradigmatic development of probabilistic approaches to object class segmentation, concerning (1) the achievements and drawbacks of the current approaches, and (2) some open issues that may concern future research in natural scene seg-

mentation.

5.1 Conceptual and Empirical Issues of Current Models

5.1.1 Performance issues

5.1.1.1 Perceptual accuracy of object boundary alignment The most important contributions of the probabilistic approaches are due to the application of semantic properties of visual categories to narrowing the gap between machine segmentation and human interpretation of a scene. One distinct aspect of this achievement is the improvement in object boundary alignment, i.e., how closely segmentation boundaries follow object boundaries. Oftentimes, semantic knowledge is required for visual interpretation of an apparently simple scene; see, for example, Figure 12.

Segmentation of semantic categories is achieved with varying degree of success. All reported works use visual inspection as the common means of performance evaluation and comparison; for selected results, see Figure 16 in the appendix. Inspection of sample output allows direct evaluation of the perceptual accuracy of the segmentation, but limited by the size of the reported cases – usually fewer than 10 test cases.

Quantitative measures of statistical accuracy are also adopted, subject to the availability of segmented test data. Accuracy/error rates are the common quantitative metrics, which measure the 'correctness' of segmentation in terms of the percentage of pixels being classified correctly or incorrectly. As mentioned in Table 2 in the appendix, many of these approaches achieve levels

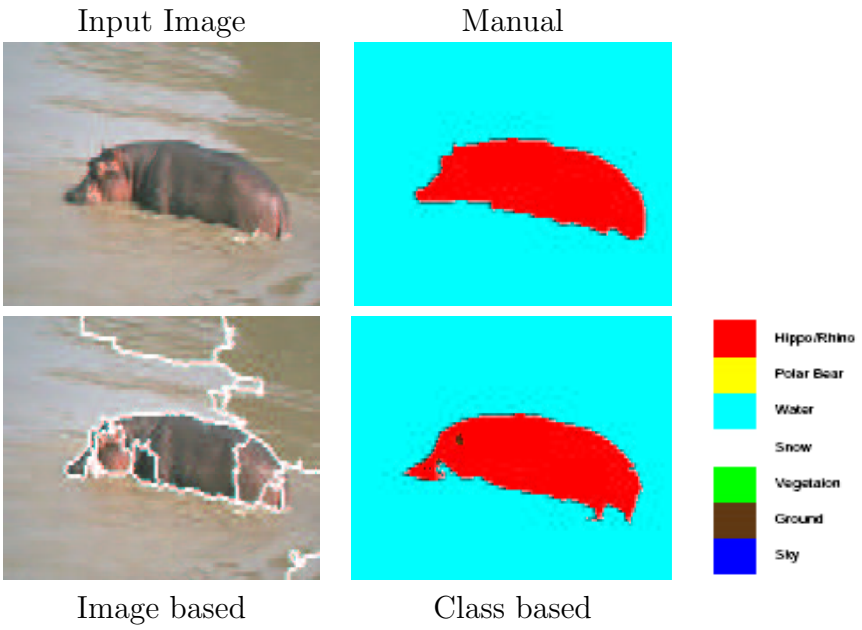


Figure 12: An example of class-based segmentation: an illustration of object boundary alignment. The class based segmentation is produced by MoCRF and the image based segmentation by mean shift. The image is labeled in terms of 7 known visual categories (the legend is shown on the far right). Source: [55].

of accuracy between the high eighties and high nineties percent. While these statistics provide a convenient summary of experimental findings over all test cases, they must be considered with great caution. These quantitative measures of correctness or accuracy of segmentation vary with the choice of training and test data. It is common to test reported work on in-house datasets or a particular subset of a much larger, publicly available dataset. These test data are usually compiled or selected to demonstrate what the tested methods are capable of. It is however unclear about how the methods perform under

increasingly less advantageous conditions. It is also a challenge to determine from a few reported cases when the methods are applicable, apart from judging by their assumptions and theoretical set-up [106, 128].

5.1.1.2 What is a "good" segmentation'? Perceptual accuracy of segmentation is an elusive concept that deserves some thought. The very idea of accuracy of segmentation assumes an image partition that corresponds to human/biological perception. This is particularly true for applications where these results are provided for human interpretation. Segmentation performed by a human expert yields the 'ground truth'. In some special cases, visual perception may be more constrained by the common purposes of expert domains. In more general situations, people perceive the world differently, across individuals, across time, across conditions and across tasks. The visual task at hand is an important factor that determines how one thinks about the 'goodness' of visual segregation.

In contrast to many other segmentation approaches, especially image-based ones, probabilistic approaches strongly couple segmentation with visual recognition. Given the task, there remains a lack of indisputable definition of 'correct' segmentation of a scene. Figure 13 displays the segmentation of each of the three images (in the leftmost column) by different individuals; for further details, see [104]. Different individuals carve up an image in distinctly different ways[8], resulting in not only variations in details, but also qualitative differ-

[8]Some may observe from these results consistent organization in human perception [105, 104]. This does not alter the fact that people do not completely agree among each other on

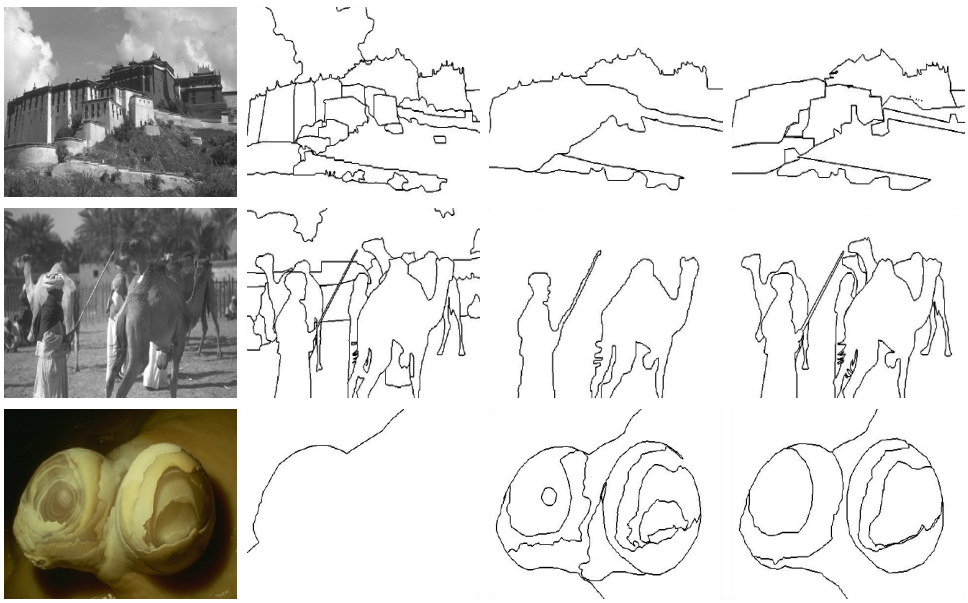


Figure 13: Manual segmentation of a scene. Segmentation of the images in the first column from the left by different individuals is displayed on the right. Each image is segmented by 3 subjects selected from a group of 10 undergraduate students under the instruction: *Divide each image into pieces where each piece represents a distinguished thing in the image. It is important that all of the pieces have approximately equal importance. The number of things in each image is up to you. Some between* 2 *and* 20 *should be reasonable for any of our images* [106]. Source: [105, 104].

ences. Each of these segmentation results, expressing human perception of an image, claims to be the ground truth of visual segmentation. As nontrivially different as they are, statistical accuracy measures machine output against a shifting standard.

Furthermore, object-embedding subimages are not equally significant; they

what features are so important that they deserve being marked out.

acquire their significance from the visual needs of observers. In many cases, this can be possible only in the task context of visual performance. In view of the ambiguities concerning the perceptual relevancy and accuracy of visual segmentation, one should take due caution on using and interpreting the quantitative measures of accuracy.

5.1.1.3 Evaluation of segmentation quality Even without this conceptual problem, one is left uncertain about how to measure perceptual accuracy in terms of statistical summary, e.g., accuracy/error rate. Due to the nonlinearity between these accuracy measures and perceptual 'goodness' of segmentation, significant improvement in perceptual quality of visual segmentation may result in only a slight increase in the accuracy rate. In other words, a small subset of pixels yields nontrivial perceptual cues for the salient and discriminative aspects of a semantic object, as illustrated in Figure 14. Despite a small difference ($< 2\%$) in pixel-based accuracy, the foreground segment of the map on the right makes it easy to conjecture a cow (or at least an animal) while it is quite hard to decipher what exactly the other segmentation results represent. This perceptual difference is made by small subimages that capture its legs and its head. The example highlights the problem of evaluating and comparing segmentation results based on pixel-based accuracy rates. Neither the qualitative nor quantitative aspects of segmentation evaluation have been addressed in the literature.

Furthermore, a scene segregation system is characterized by many functional properties – including but not restricted to correctness/accuracy, efficiency,

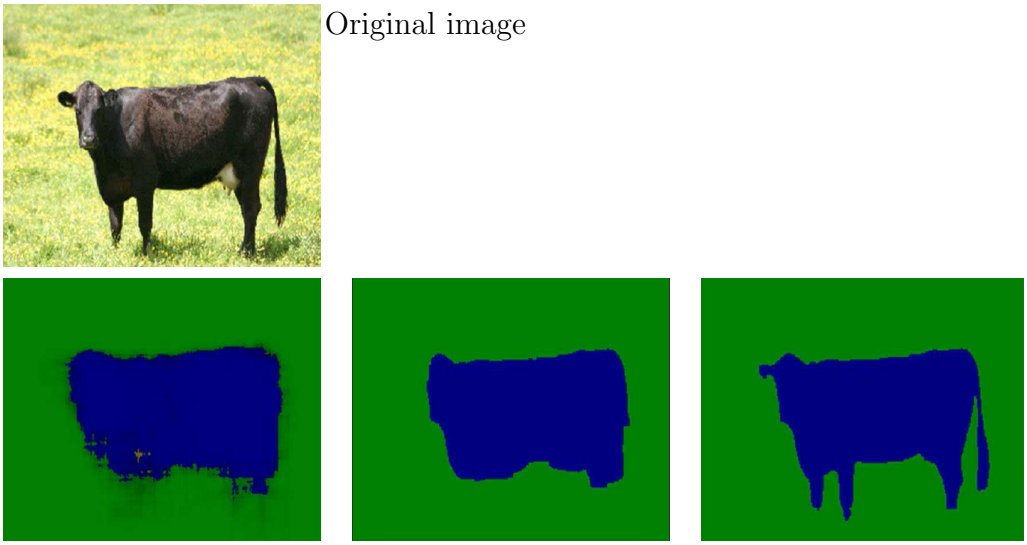


Figure 14: Statistical vs. perceptual accuracy of segmentation: an illustration. Top row: the original image. Bottom row: segmentation maps with the corresponding pixel-based accuracy rates indicated at the bottom. The class labeling is assigned using the method of textonBoost. The accuracy is measured in terms of percentage of pixels assigned to the correct class label. Source: [149].

reliability, stability, capacity, adaptivity, generalizability and robustness. A high-performance system usually presents a reasonable tradeoff between these requirements. Except for accuracy and to some extent efficiency, there is little focus on the other aspects of performance in the current literature on object class segmentation. The paucity of research on these other performance properties makes it very difficult to assess the relative merits of these algorithms in real-life applications. It is a challenge to thoroughly study these properties without a consistent and theoretically sound framework for measuring algo-

rithmic performance.

Concerning only the accuracy of scene segregation with respect to object boundary alignment – the focus of research in the current literature – there are major issues to be addressed. At present, most of the data sets used for semantic-based visual segmentation or visual understanding experiments are compiled from online or proprietary image collections with inadequate control over the range of visual conditions for experimental purposes. A more elaborate experimental framework is therefore required to assess the performance characteristics of an algorithm. More controlled experiments with specific scene and stimulus conditions may be instrumental in probing into specific performance properties. A systematic investigation of performance properties may require a mixture of synthetic and natural image data. An important component of this framework is the methodology to guide the construction of datasets. The framework should always allow comparative studies to explore in a systematic way important characteristics of major approaches under rigorous experimental control with respect to the full range of visual conditions pertinent to the tested methods. No sound methodology of performance evaluation/comparison can be complete without representative metrics of performance measurement that can capture the perceptual accuracy of segmentation.

5.1.2 Issues pertinent to categorization in visual segregation

5.1.2.1 Inadequacy of object formation The progress towards the goal of visual segregation of semantic categories should not be overstated. In the lit-

erature, the concept of object segmentation is usually narrowly and sometimes arbitrarily defined. Many approaches [75, 103, 54, 55, 149, 135, 18], in particular but not restricted to those aiming at scene segmentation, entertain no concept of an object. They fail to segregate object instances from one another, resulting in grouping multiple instances, especially those spatially contiguous, into a common perceptual unit (also see Figure 16 as well as Table 2 in the appendix). This does not create a serious problem for the specific task of scene segmentation which is not concerned with object-level segregation.

Yet, the lack of unit formation in their perceptual apparatus leads to a level of granularity that is inappropriate for many visual recognition tasks. For instance, the subimages carved out from the image in column (a) and (b) of Figure 15 clearly suggest the presence of a motorcycle or a four-legged animal. In contrast, those in column (c) and (d) do not provide very helpful cues for determining the objects in the scenes. Similarly, the segmentation in column (c) and (d) is also inadequate for many other visual tasks, such as object-based image coding and editing.

Much of the discriminative power of scene segmentation and other approaches springs from the classifiers used to capture image statistics of local features. With little reliance on abstract organization of local features, many of these approaches are applied to long or medium shot, outdoor scenes describable by coarse-grained, amorphous categories, such as sky, water, vegetation, and others. These categories are usually characterized by a small number of simple features but not by any definite form or structures. It is a challenge for these

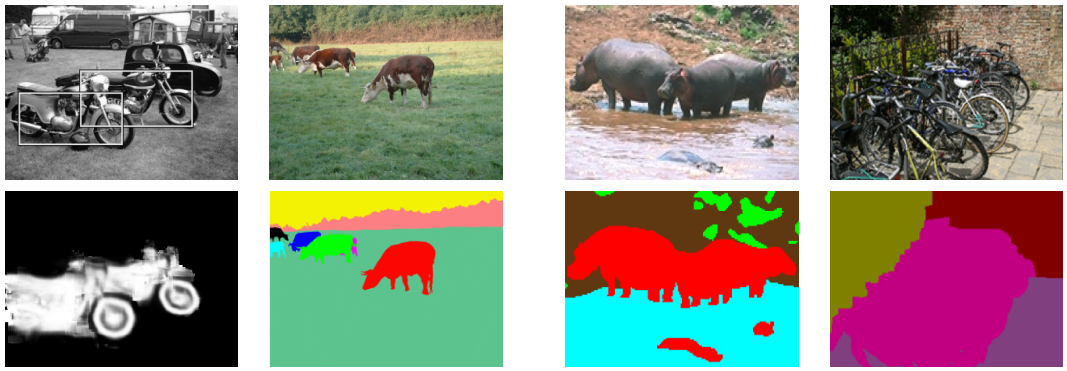


Figure 15: The role of unit formation in object perception: an illustration. Top row: the original images. Top: original images. Bottom: segmentation (manual or machine segmentation/ labeling). Individual object instances are segmented in the images in column (a) and (b) but not in (c) and (d).
Source: [92] (machine output), [52] (manual), [54] (manual), and [149] (machine output) in the order of presentation from left to right.

loose assemblies of local features to segregate object instances that require sufficient ability to discriminate between fine-grained details organized in highly specific structures.

The class of foreground/background segregation approaches [15, 89, 17, 14, 92, 80, 16, 79, 88, 94, 172] aims at more fine-grained distinctions between target objects and their background. Instead of loose assemblies of local features, models of spatial structures, usually with respect to some object-centered frames, provide important advantages to capture the recurring properties of target object classes. Constellations of local features, in particular, yield robust representations for object differentiation under partial occlusion conditions; see [89, 92, 88] for instance.

Yet, the problem of foreground/background segregation is commonly viewed as a binary partition problem (or its variants), dealing with one particular visual category, e.g., cows, horses, cars, bikes or others, and 'pushing' all other objects into the background, even though some of these other objects may stand in front of the known object (for examples, see Figure 16 in the appendix.) It is debatable whether this class of segmentation can be described in a strict sense as foreground/ background segregation. This idea of 'foreground' is thus defined not in terms of the role and saliency of subimages, but rather according to what a system 'knows' about the world. Moreover, many of these approaches pull out only a single instance from a scene. The performance is often studied with a test set, where each image contains a single and usually close-up shot of an instance. In other words, this class of algorithms segregates only a single individual instance of a specific object class instead of carving up an image along semantic (object class) boundaries. It is the latter that scene segregation ultimately seeks. At least in theory, a simple extension to the existing, binary classification paradigm of foreground/background segregation is possible, by extracting the instances recursively one class at a time. In practice, apart from the computational time required for completing segmentation, how badly the system performance degrades with the introduction of each additional visual category is an open question.

The problems of representation express themselves in a more fundamental way than simply in terms of the number of categories allowed for segregation. As a result of the underdeveloped representational architectures of the prevailing

models, object classes may not necessarily be defined in the literature in a way that corresponds to human intuitive perception, but rather formed with respect to the discriminative capability of data-driven classifiers. It has been suggested that buildings, for instance, are not a 'good' class since they come in many styles and appearance; rather 'stonework' is a better choice due to its simple and 'homogeneous regular texture' for discrimination [75].

5.1.2.2 Perceptual constancy It is hard to overstate the importance of perceptual constancy of representation. A stable perception of semantic objects is not possible without some level of constancy that maintains stable and robust representation despite the myriad of their appearances caused by extraneous conditions, such as viewing distance and direction, illumination, occlusion, and others. Despite much research effort expended on the class-wise variation due to the appearance differences across member objects, the issues of representational constancy of objects in visual categorization remain largely unexplored in the literature.

It is rare for existing models of object class segregation to consider more than a specific view of object instances at a particular scale. A limited extension of these view specific approaches is to infer the object embedding by flipping the view specific prototypes along the horizontal axis [164]. The 3DLCCRF [60] uses a decision forest to model a multiple appearance models for car parts from four different viewpoints, one for each 45° viewing range. It is able to recover object instances from different points of view, but with 'prohibitive computational time' and rather weak object boundary alignment. This however

represents a rudimentary attempt to address a very important but seriously under-explored issue in visual categorization.

Similarly, there is no explicit representational scheme to encode scale dependent appearances of object instances. The term 'multiscale' is adopted in the literature as usually referring to the level of granularity, i.e., the size of neighborhood from which local classifiers pool information to compute their responses. The issues of scale are sometimes addressed by resizing the prototypes in the visual vocabulary to a small number of fixed scales [60]. This is hardly a general solution to the problems of scale constancy of visual representation. The advantages of scale space representation which successively reduces minute details and emphasizes salient perceptual structures, have not been systematically exploited[9].

The formation of a scale insensitive perception requires more than a scale-space representation. The appearance indeed changes as the viewing distance varies. The canopy of a forest may be easily recognized from a distance by some regular texture, which is completely insufficient for a close up view where the tree trunks, branches and leaves become salient in the scene. Similar to their viewpoint counterparts, scale insensitive representations have to take into account the emergence and disappearance of saliency features as they move towards or away from the observer. Visual representation with perceptual constancy is a yet to be explored area of research. Until these issues are properly addressed,

[9]The exceptions include the approaches of segmentation based on adaptive segregation of saliency regions into a hierarchical structure of scale varying representation [14].

'object class segmentation' falls short of the claim of segregating objects.

5.1.2.3 The issues of representation for abstract visual concepts The prevailing models of feature-based classification lack adequate mechanisms of organization and abstraction that are essential for developing informative and parsimonious representations of complex visual concepts. There are certainly circumstances, especially those affording strong[10] *a priori* knowledge of object classes, that a simple conjunction of a few image features is sufficient for classification. Certain object classes, under the assumption of simple scene conditions, can be satisfactorily recognized or classified by some features or parts without considering the full complexity of the formation of an object as a whole [6]. These circumstances may not be rare but they constitute only limited cases among the manifold and diverse energy patterns that can be emanated in the natural, visual environment.

Yet, both theoretical and empirical studies have emerged from a broad spectrum of vision research, suggesting the essential roles of complex pattern integration and perceptual abstraction in semantic-based visual tasks [36, 5, 6, 108, 83, 7, 137]. It is questionable how far the current models can be extended beyond those limited cases, using simple mappings of primary image features to semantic classes without introducing the necessary organization and abstraction of perceptual patterns.

[10] These are the circumstances where both the variability of visual appearance are low and the number of expected object classes are small.

5.1.3 The issues of inference in visual segmentation Inference has been a central problem in the literature of object class segmentation. Many inference models and approximation techniques are extensively explored in the context of object class classification. Indeed, it would be no exaggeration to attribute many advances in semantic-based segmentation to the recently developed ideas and techniques of probabilistic inference that have been applied to the problems.

At present, however, the capacity of many existing approaches is indeed greatly constrained by the complexity of inference. Stochastic search approaches, such as image parsing [157], remain very expensive in sampling the solution space, even though the scene is modeled in terms of few and simple object classes with little variations. Many object classification algorithms for foreground/background segregation are also expensive. Few authors explicitly mention the computational cost of their segmentation, but there are reports on experiments that take prohibitive amount of time in inference – see [165, 60] for instance.

The complexity problem of inference is not restricted to specific algorithms, but has general implications for probabilistic approaches to visual segmentation across the board. A probabilistic model is essentially a system of normalized energy defined over a solution space. As mentioned, the partition function contains all the information over all possible states – a space which increases combinatorially with the number of visual categories. Approximation solutions are the only practical options for many inference problems in object

class segregation due to their computational complexity, including many message passing algorithms developed for non-tree structured graphical models. Many of these approximations help keep the cost feasible but it may rise fast as the states proliferate with the number of visual categories. How well the underlying correlation between explanatory factors can be encoded is limited by the approximation assumptions leading to lower accuracy and, in some cases, greater instability. These problems may be further exacerbated by the increasingly complex surfaces of model energy, due to rising number of visual categories, higher variability of object class appearance, and more complex relationship between object classes in the feature space.

That inference is too computationally demanding for many visual problems remains a strong drive to develop less expensive and more flexible frameworks of visual inference. To avoid the complexity associated with normalization, there is active research in the recognition community on (non-probabilistic) energy-based models, that have been less explored in the context of object class segmentation [57, 85, 83, 84, 130]. At present, the complexity of inference techniques places stringent restrictions on the usefulness of current models. Reducing these restrictions is critical for the further development of the framework into a viable approach for natural scene segmentation in unrestricted settings.

It should be noted that inference and representation are strongly coupled in any solution to the problem of visual understanding. Without an adequately organized representation, the mapping from visual appearance into seman-

tic categories becomes highly intricate and non-linear. Working with a crude representation inadvertently places enormous computational burdens of visual interpretation completely upon the inference procedure – thereby greatly exacerbating their complexity. The prevailing approaches are indeed unified by a common theme, which seek in different ways to apply the classical segmentation criteria, region-based or edge based, to local classifier responses. Due to the high dimensionality of the feature spaces required by inter- and intra-class variations, extremely lengthy training periods are usually required for binary or multi-class labeling[11]. Alternatively, visual representation can be organized in a more expressive and flexible way to enhance the discriminative capability, therefore alleviating the burden on the inference. Apart from the benefit of lower cost, simpler inference problems demand less approximation and therefore higher accuracy and greater stability. This latter path is rarely taken.

5.1.4 The issues pertinent to the relations between segmentation and categorization The conceptual as well as computational relations between segmentation and classification/recognition are issues at the core of object class segmentation. In the prevailing models, semantic-based segmentation may be thought of as a special class of visual categorization, viz, the pixel/superpixel classification. The focus is on the issues of modeling semantic-

[11] Even scene segmentation, despite its coarse-grained representation and the lack of object segregation capability, may be computationally expensive, taking as long as $14,000$ hours to train a classifier on a 21 class training set of 276 images on a 2.1 Ghz machine with $2GB$ memory if random feature selection is not used [149].

based influences on pixel classification. Iterative processes of region growth are arguably a rudimentary form of these influences in the general framework of visual segmentation based on classification. Starting with a number of seeds selected based on classifier responses, some spreading (or refinement) processes are set in motion, which aggregate visual inputs into object class embedding regions through negotiation and competition among visual categories; see [103, 80, 79]. These approaches invert the classical order of visual processing between image-based and semantic-based perceptual processes. In their own ways, the stochastic search approaches eliminate the distinction between segmentation and recognition by merging the two processes within a single inference structure of sampling segmentation models [157].

These approaches recast the problem of segmentation in a fundamental way. In the classical paradigms, segmentation is pursued in order to delineate salient regions that correspond to significant visual events or to provide perceptual units for knowledge-guided visual analysis. Associated with this view are strong tendencies to postulate sharp distinctions between different perceptual processes and unidirectional visual pathways [100, 101, 102]. Yet, a need for more complex interactions between different visual processing areas has been recognized since the early years of vision research, in addition to the purely feedforward communication which allows only one-way signal transmission.

The focus of many earlier approaches, such as those emphasizing the cycle of representation, was placed upon the organizational architectures, many of which involve feedback and recurrent processes of visual reasoning; for fur-

ther discussion, see [65, 117, 155]. Recent research has been providing evidence for interactive processes underlying human perception; for segmentation and object segregation, see [126, 127, 160, 125, 119]. Computational models are sought to overcome the limitations resulting from these restrictions by re-introducing interaction between segmentation/grouping processes and semantic-based processes.

It can be hardly overemphasized that the discovery of the semantic-based influences on perceptual formation of objects and their segregation does not eliminate the relatively independent and indispensable functions of unit formation processes. Viewing segmentation as a classification problem, however, many current approaches to object class segmentation reduce the role of segmentation to delineating the spatial extent of object instances after they are recognized. The alternatives are those strategies that use superpixels generated by image-based partitioning techniques. Semantic-based segmentation is accomplished by classifying these superpixels in terms of membership in visual categories. The contribution of the influence of complex and abstract visual concepts is restricted to classifying preprocessed regions.

Common to these strategies, with some rare exception [71], is their feedforward approach to visual representation. Thus no means are available for different perceptual processes to collaborate in determining the significance and implications of visual features across levels of scale and abstraction. For instance, there is no way to redefine superpixels in light of their semantic-based characteristics – perhaps the simplest form of collaboration between image-based

and semantic-based segmentation processes.

These approaches may be helpful for image processing applications such as object-based edition or for providing perceptual cues for content-based analysis, but greatly reduce the role of segmentation and grouping processes in encoding visual abstraction and symbolic constructs. From the broader perspective of visual analysis, segmentation is nothing less than an integral part of pattern organization and visual abstraction in the very formation of perception. When scene segregation is viewed as an integral part of object perception formation, no computational account of the phenomena would be sufficient without considering the interconnections between different processes that subserve perceptual representation and segregation. Indeed, natural scene segregation provides a unique opportunity to explore these important aspects of computational perception; and yet they are in general overlooked in the literature of image segmentation

5.1.5 The issues pertinent to the systematic properties Lacking so far is a general, theoretical approach to the issues concerned with the capacity, generalizablity, robustness and adaptivity of visual categorization processes in the probabilistic approaches to categorization/recognition and visual segregation. These functional properties together with view/scale constancy are referred to collectively as systematic properties or functions. There is no dispute over the importance of these properties. However, it is nontrivial to address these issues under the prevailing paradigm of visual categorization and segmentation. One cause of the problems is the theoretical structure of

the current approaches to visual categorization – the structure which makes it very difficult to deal with the systematic issues in a general way.

As discussed, the categorization framework is built with a shallow structure where image features are extracted and then classified into visual categories through a trained probabilistic model. Therefore, few options are available to address the systematic issues other than looking for a rich feature space so that probabilistic inference is afforded a distinctive representation to describe a scene in terms of a sufficiently large number of complex categories. In principle, the task would become more tractable if features associated with different categories can be separated from each other by a relatively sparse background with discernible margins. To this end, a feature space is sought to project the characteristic visual properties of the categories onto a finite number of compact regions. This rich feature space is in general of very high dimension. High dimensionality certainly comes with a cost in terms of computational complexity.

The complexity of visual categorization arises from a visual environment which demands highly organized processes of perceptual integration and abstraction to unravel the physical energy patterns in terms of perceptual concepts. In the absence of adequate support of functional organization, other than a few simple layers of feature-extraction and classification, visual categorization is torn between, on the one hand, the high dimensionality of a rich feature space and the complexity of classification models sufficient for the systematic properties of the system and, on the other hand, the simplifying assumptions of approx-

imation and computational heuristics required for probabilistic inference. In practice, computational feasibility takes priority over relevancy.

Obvious from these observations are two classes of solutions for further exploration. It should be noted that despite their different approaches to the problem, these solutions are not necessarily exclusive and competing. In principle, the systematic issues can be addressed by developing effective inference structures that are adequate for complex probabilistic models defined over a rich feature space without undue approximation restrictions.

The second class of solutions emphasizes the central ideas of perception formation and the importance of its organization. In particular, the representational capacity and the related systematic properties of a visual system are by and large determined by its own organizational structure, whereby the content of visual stimuli can be unraveled, abstracted and reorganized for better discriminability and easy read-out. Indeed, studies in pattern classification and pattern theory make similar observations that pattern models with deeply-structured and well-organized architecture are capable of discerning the order of complex data patterns through interactive stages of pattern transformation; see [83, 84, 7, 51]. Evidence from vision research also attests the essential role of recurrent stages of perception formation in visual categorization. In particular, perceptual groupings, contours, surfaces and objects are important perceptual structures in organizing visual representations for categorization as well as recognition.

5.2 Summary

This paper provides a conceptual account of the theoretical structures that drive the current advances in natural scene segmentation. These structures are elucidated in terms of the guiding ideas and computational principles of image segmentation. A strong theme in this development is the view of image segmentation as an inference over image representations. These representations are models of visual patterns that capture the characteristics of visual coherence pertinent to the interpretative criteria of segmentation based on discriminative clues afforded by the visual appearance of subimages (regions) and their boundaries. Image-based segmentation may be concerned with some homogeneous or smooth intensity patches or piecewise smooth contours. The relevant classes of visual patterns proliferate as increasingly richer semantics are introduced to capture the underlying structures in visual appearance of perceptual categories that are highly variable over some complex measurement spaces. This leads to a fusion of image segmentation and pattern recognition – an approach that has pioneered clustering/classifier-based segmentation and has been increasingly adapted in many semantic-based precursors as mentioned in the previous discussion.

To understand complex scene composition in terms of semantic categories that rely on conceptual distinctions rather than visual differences, visual patterns can no longer be easily and unambiguously mapped to a scene configuration in the solution space. It is necessary to introduce semantic explanation and abstraction to bridge the semantic gap. To capture the richer semantics of seg-

mentation, an adequate method must be equipped with sufficient structures of representation and inference to incorporate a wide range of discriminative criteria and to integrate computation of diverse recognition and segmentation processes across different layers of abstraction. To cope with the complexity of scene composition, it would be more feasible to model an image as stochastic processes in terms of some underlying statistical structures and plausible interpretations. This provides strong motivation for viewing image segmentation as an optimization problem seeking the best interpretation over all plausible ones. Energy based representation in its various formulations provides an efficient and well-studied framework for modeling observation, image semantics and segmentation constraints.

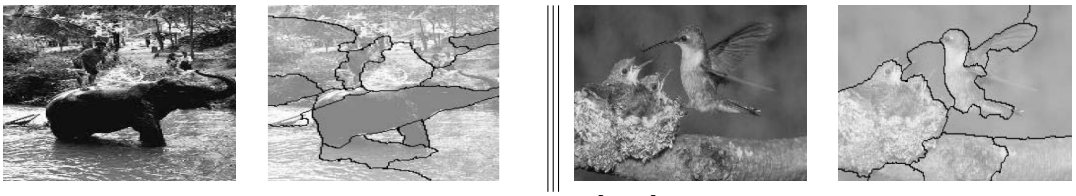
The bulk of the paper discusses the recent developments of probabilistic approaches, which seek object-embedding subimages enclosed by object-specific boundaries. With the benefit of the recent advances in semantic-based visual analysis on one hand and those in probabilistic modeling and inference on the other hand, these methods attempt to forge bridges between the semantic properties of natural scenes and visual features discovered by a set of object-specific probes, i.e., filters/classifiers. From a historical perspective, these approaches may be viewed as applying the basic principles of image-based segmentation to parsing an image into meaningful categories based on content-specific visual patterns.

At present, there are many open questions that should be addressed in future research. The performance of existing approaches are greatly constrained by

the complexity of inference techniques and more efficient procedures are required. There is a pressing need to devise effective methods for investigating the functional/performance properties of segmentation systems. The existing approaches fail to adequately provide organized representations to capture the complex relationships between image appearance and visual categories under general conditions of natural scene segmentation. Figure/ground segregation of individual objects remains an unsolved issue in most of the current frameworks. The problems pertinent to perceptual constancy and the system-level organizational principles of natural scene perception have rarely been addressed in the literature. The functional roles of semantic-based segmentation in the context of object perception should be clarified in a more theoretical and systematic manner. Context modulation of scene segmentation by other computational processes, in particular, those responsible for non-local, complex and abstract object perception have yet to become a focal area of inquiry.

Appendix

This appendix presents supplementary materials in support of the discussion of the current developments in the text, and in particular, the remarks presented in Section 5. All algorithms discussed in this paper use qualitative evaluation by visual inspection of segmentation results for assessing and comparing performance. A selected set of reported segmentation results is presented in Figure 16, which may provide some ideas of the performance of representative algorithms. Also included in this appendix are summaries of the algorithms. Table 1 which highlights the algorithmic features and theoretical properties of the representative probabilistic approaches, and Table 2 summarizes the major aspects of empirical experiments used for performance evaluation.



Ren et al. [134]

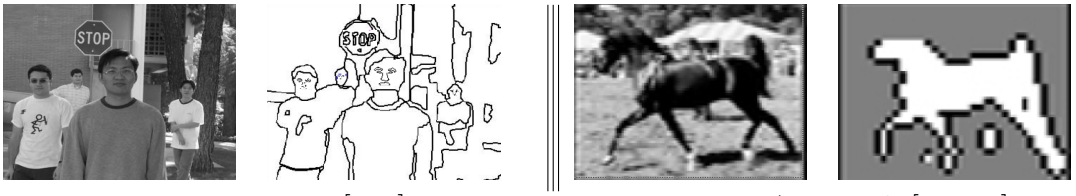
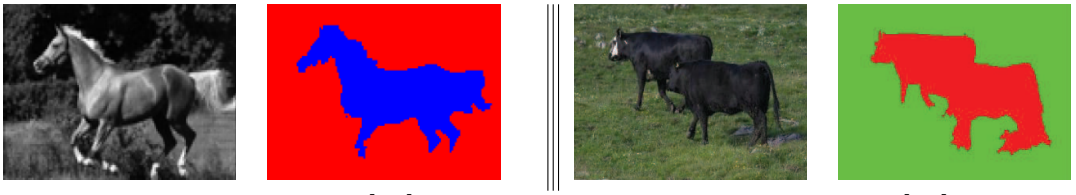
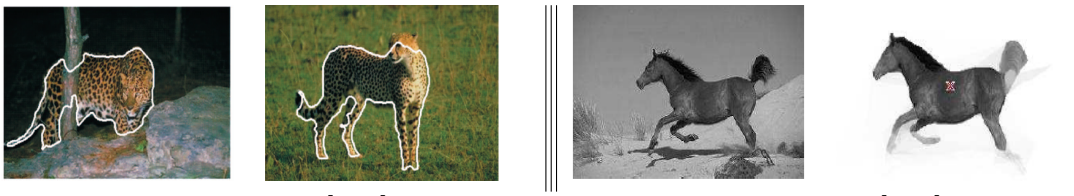


Image parsing [157] Jigsaw Approach [15, 17]



Borenstein et al. [14] Spatial-LTM[23]

Implicit shape model (ISM) [89, 92, 88]



Zöller et al. [172] Ren et al. [132]

Figure 16: Performance of object-class segmentation: results of selected approaches.

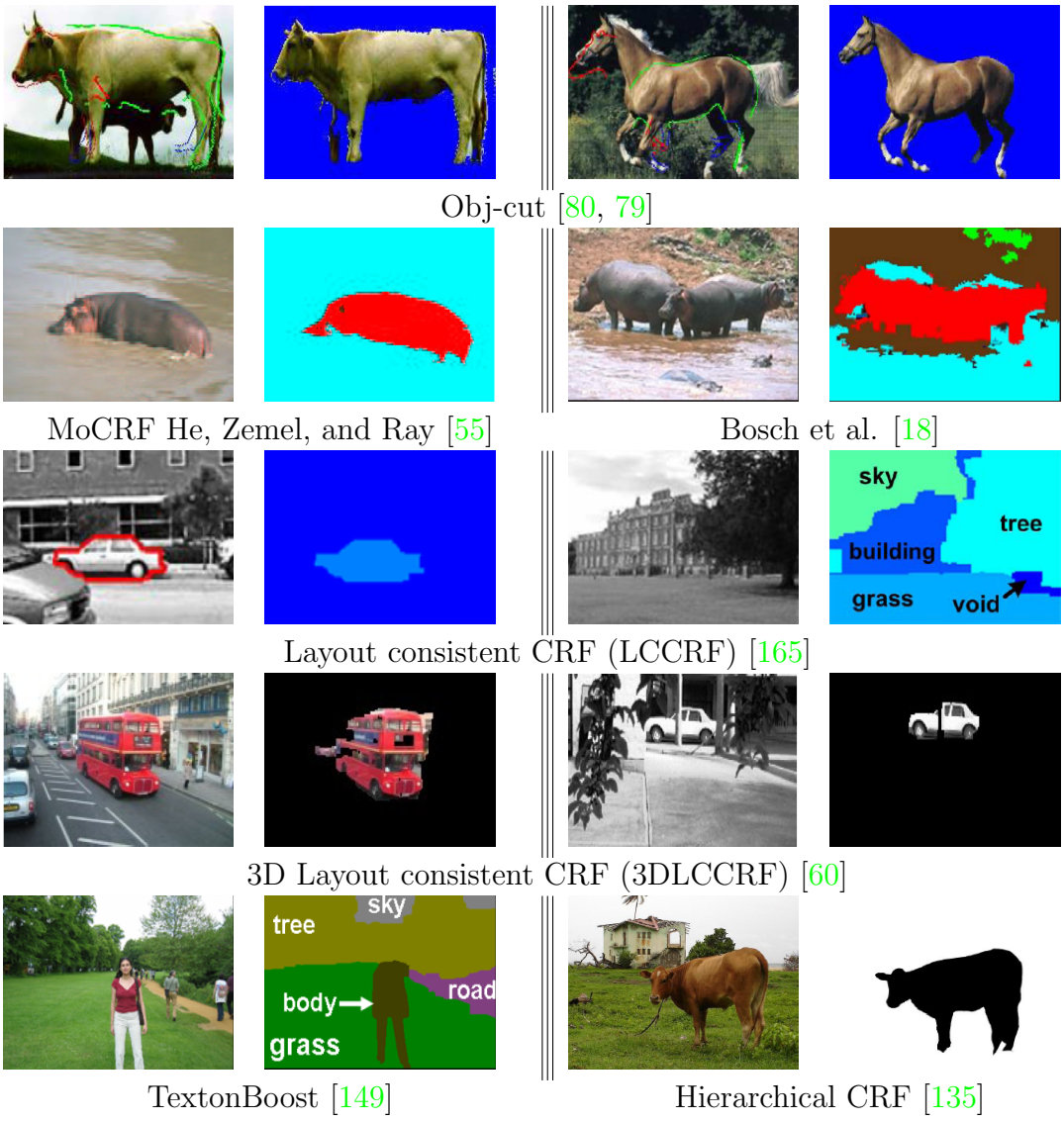


Figure 16: Performance of object-class segmentation (contd): results of selected approaches.

Table 1: Algorithmic features of selected approaches. A note of explanation is found at the end of the table.

Algorithms	Approaches	Occlusion	View	Remarks
Image parsing [157]	Stochastic search; Parsing graph; Discriminative/ generative model integration.	×	specific	Theoretically well grounded; Online determination of number of segments; Non-trivial to design irreducible chain; Combining image- and class-based features; Assuming objects with little variability; Weak results; Slow.
Ren et al. [134]	Stochastic search; merging image-based segments with object class grouping cues.	∼	∼	Combine image-based segmentation with object class cues; fragment individual object; expensive; no object labels.
Jigsaw [15, 17]	Use a codebook of visual patterns	×	specific	Pioneer approach; Simple; Not suitable for classes with great variability; Huge codebook; Segment single instance

Continued . . .

Table 1 – continued

Algorithms	Approaches	Occlusion	View	Remarks
Borenstein et al. [14]	Combine jigsaw approaches and image-based visual processing cross scales.	×	specific	May improve on the boundary alignment; Not able to segment complex objects; Need huge codebook for variable classes.
Levin et al. [94]	Codebook of visual patterns; CRF; Combine object class and image-based cues.	×	specific	Feature selection and image-based cues help to reduce the size of codebook patterns; Not suitable for complex objects with highly variable forms and appearances.
ISM [89, 92, 88]	Use codebook of visual patterns; Configuration modeling.	√	specific	Allow multiple instances; Not suitable for classes with great variability in forms; Vocabulary size is enormous for complex objects.

Continued . . .

Table 1 – continued

Algorithms	Approaches	Occlusion	View	Remarks
Kokkinos et al. [71]	Use codebook of visual patterns; Morphable model	limited	specific	Allow semantic influence on segmentation modification; Not suitable for classes with great variability in forms; llimited representation capacity.
Martí et al. [103]	Interleave recognition and scene labeling; Graph-based models of semantic relations of object classes; Use fuzzy reasoning.	∼	∼	Need very specific models for target scenes; Boundary alignment may be affected by over-segmentation; No instance differentiation; More suitable for visual categories with little fine details; Limited capacity of representation.

Continued . . .

Table 1 – continued

Algorithms	Approaches	Occlusion	View	Remarks
Todorovic et al. [153]	Subimage-tree based representation, tree matching	*limited*	∼	Too complex except for a very small training set; rotation invariant but week in scale and view invariance; Weak segmentation with missing parts; Not situable for complex class with high variability or more than a few salient parts; Low representation capacity; Very expensive.
Konishi et al. [75]	Scene labeling; Conditional probability model of labeling given classifier responses	∼	∼	Simple; Fast; No instance differentiation; Not suitable for visual categories with complex and variable appearance; Very limited representation capacity.

Continued . . .

Table 1 – continued

Algorithms	Approaches	Occlusion	View	Remarks
Russell et al. [140]	Latent topic discovery model; superpixel classification.	∼	∼	Weak object boundary alignment; Uncertain performance with visual categories with complex and variable appearance; Limited capacity of representing general classes.
Spatial-LTM [23]	Latent topic discovery model; superpixel classification.	∼	∼	Experiments include only unoccluded objects; No evidence for instance differentiation; Some view tolerance; Uncertain performance with visual categories with complex and variable appearance; Limited capacity of representing general classes.

Continued . . .

Table 1 – continued

Algorithms	Approaches	Occlusion	View	Remarks
Markov Field Aspect Models [162]	Latent topic discovery model; Spatial coherence constraints via random fields	∼	∼	Coarse region classification; poor object boundary alignment; Uncertain performance with visual categories with complex and variable appearance; Limited capacity of representing general classes.
Obj-cut [78, 80, 79]	Energy model (GRF†); Image-based and object class integration; Part configuration.	self-occlusion	specific	Able to segregate articulated objects; time consuming for training and segmentation; Assume rigid objects; Segregate a single instance.
mCRF [54]	Energy Model (CRF); Scene labeling; Multiple level of scales	∼	∼	Not able to segregate individual object instances ; Need training set of substantial size; expensive; Better performance on coarse scale image with relative little fine details

Continued . . .

Table 1 – continued

Algorithms	Approaches	Occlusion	View	Remarks
MoCRF [55]	Energy Model (CRF); Scene labeling; Context constraints.	∼	∼	Not able to segregate individual object instances ; Need training set of substantial size (thus considerable training time) ; Better performance on coarse scale image with relative little fine details
Ren et al. [132]	Energy model (CRF)	×	specific	Use cues from different scale of analysis; More suitable for simple object class with little variability in form and appearance; The representation capacity is limited.

Continued . . .

Table 1 – continued

Algorithms	Approaches	Occlusion	View	Remarks
LCCRF [165]	Energy model (CRF); layout consistent modeling.	√	specific	Model consistent configuration for occlusion explicitly; Allow multiple instances; Not able to recover disconnected parts of individual instance; Not suitable for complex classes with high variability in forms and appearance; Low representation capacity.
3DLCCRF [60]	Energy model (CRF); layout consistent modeling; multiple view modeling.	√	multiple	Layout model; Allow multiple instances; Some scale tolerant; Very expensive; Not suitable for complex classes with high variability in forms and appearance; Low representation capacity.

Continued . . .

Table 1 – continued

Algorithms	Approaches	Occlusion	View	Remarks
TextonBoost [149]	Energy model (CRF); Scene labeling.	∼	∼	Incorporate longer range information, incl. spatial and contextual inf.; Extremely expensive; Not instance differentiation; Not suitable for complex classes with variable forms and appearance; Low representation capacity.
Hierarchical CRF [135]	Energy model (CRF); Tree-structured model of contextual constraints over multiple scales; Combine image-based and object class cues.	∼	specific	Use of superpixels and multiple scale analysis improves instance boundary alignment; Tree structured model speeds up inference; Remain expensive for training the model; No instance differentiation based on object concepts; Limited capacity of representation.

Continued ...

Table 1 – continued

Algorithms	Approaches	Occlusion	View	Remarks
Bosch et al. [18]	Scene labeling; Region growth according to homogeneity and contrast measure; Seeded by object class classification.	∼	∼	No instance differentiation; Not suitable for complex object with refined details and variable appearance; Limited representation capacity;
Zöller et al. [172]	combine clustering with semantic map; shape information.	tolerant	specific	Not suitable for complex object with refined details and variable appearance; Weak object boundary alignment; Limited representation capacity;

† **Notes of explanation**

∼	Not applicable
×	Assume no occlusion.
√	Assume occlusion.
CRF	Conditional random field
GRF	Generative random field
LCCRF	Layout consistent conditional random field
Occlusion tolerant	Segmentation the object without separating it from the occluding foreground

Table 2: Performance and evaluation. Number of object classes refers to the maximum number of classes allowed for a particular training sets.

Algorithms	Type of training set	Number of classes	Testing classes	Evaluation methods	Remarks
Image parsing [157]	unlabeled	2	faces (frontal) + text	Visual	
Ren et al. [134]	labeled image	1	bear; tigers; lions, deers, leopards, elephants, birds, horses, penguins, humans, landscapes, buildings, humans	visual	
Jigsaw [15, 17]	manual	1	horses	visual & accuracy	88 - 95 % pixel based accuracy
Borenstein et al. [14]	manual	1	horses	visual	
Levin et al. [94]	manual	1	horses	visual & error percentage	4.4 to 8% pixel based error rate; using 4 to 10 patterns.
ISM [89, 92, 88]	manual	1	cars; bikes;cows	visual	
Kokkinos et al. [71]	manual	2	cars, face	visual	

Continued . . .

Table 2 – continued

Algorithms	Type of training set	Number of classes	Testing classes	Evaluation methods	Remarks
Russell et al. [140]	labeled or non-labeld	4 - 23	bicycles, cars, signs windows, buildings, roads, sky, and other unspecified objects	visual and accuracy	Interaction ratios vary with categories ranging from 0.09 to 0.77
Spatial-LTM [23]	labeled or non-labeld	1 - 28	horses; cows; airplanes, cars, brads, motorbikes, faces;faces, leopards, motorbikes, bonsais, trains, cougar faces, crabs, cups, dalmatians, dolphins, elephants, euphoniums, ewers, ferries, flamingos, grand pianos, Joshua trees, kangaroos, laptops, lotuses, schooners, soccer balls, starfishes, stop signs, sunflowers, watches, yin yang	visual and accuracy	Accuracy rates vary with categories, ranging from low 50% to high 90% with average accuracy rates at 67%

Continued . . .

Table 2 – continued

Algorithms	Type of training set	Number of classes	Testing classes	Evaluation methods	Remarks
Markov Field Aspect Models [162]	labeled	13 - 21	buildings, grass, trees, cows, sheeps, sky, aeroplanes, water, faces, cars, bicycles, flowers,signs, birds, books, charis, road, cat, dogs, bodies, boats	visual and accuracy	Average patch-level accuracy varies with categories ranging from high 70% to low 80%
Martí et al. [103]	manual	4	leaves, roads, sky, ground	visual & accuracy	87% pixel based accuracy; test on very limited set of image data.
Todorovic et al. [153]	unlabeled	1	faces, cars	visual	
Konishi et al. [75]	manual	5	vegetation, air, building, roads, cars	visual & classification errors	Classification error varies with visual categories.

Continued ...

Table 2 – continued

Algorithms	Type of training set	Number of classes	Testing classes	Evaluation methods	Remarks
Obj-cut [80, 79]	labeled images	1	cows; horses	visual & accuracy	89.39 - 99.53% pixel based accuracy rate for a limited subset of training data.
mCRF [54]	label images	7	hippopotamuses, polar bears, water, snow, vegetation, ground, sky, road marking, road surface, street objects, cars	visual	
MoCRF [55]	labeled images	7-11	hippopotamuses, polar bears, horses, tigers, wolves/leopards, water, snow, vegetation, ground, sky, road marks, road surface, street objects, cars, fences	visual	
Ren et al. [132]	manual	1	horses	visual; precision & recall.	Shape and gestalt cues are important for performance.

Continued . . .

Table 2 – continued

Algorithms	Type of training set	Number of classes	Testing classes	Evaluation methods	Remarks
LCCRF [165]	manual	1 - 3	cars; faces; building, trees, sky, grass	visual; accuracy	96.5% of pixel based accuracy and 67% of instance-based accuracy.
3DLCCRF [60]	manual	1	cars	visual; accuracy	77% of instance-based accuracy.
TextonBoost [149]	manual	1 - 5	building; grass; tree; cow; sheep; sky; planes; water; faces; cars; bikes; flowers; signs; birds; books; chairs; roads; cats; dog; bodies; boats	visual; accuracy	$74.6 - 88.6\%$ of pixel-based accuracy.
Hierarchical CRF [135]	manual	1	cows, cats, cars	visual; precision & recall	Performance are contributed mainly by classifiers instead inference over the random field

Continued . . .

Table 2 – continued

Algorithms	Type of training set	Number of classes	Testing classes	Evaluation methods	Remarks
Bosch et al. [18]	manual	5 - 7	rhino, polar bear, sky, grass, roads, vegetation, land, snow, water, ground	visual & accuracy	86.76% of accuracy.
Zöller et al. [172]	manual	1	wild cats	visual	

References

[1] Narendra Ahuja. A transform for multiscale image segmentation by integrated edge and region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(12):1211–1235, 1996.

[2] Ruzena Bajcsy, Franc Solina, and Alok Gupta. Segmentation versus object representation −− are they separable? In *Analysis and interpretation of range images*, pages 207–223. Springer-Verlag New York, Inc., New York, NY, USA, 1990.

[3] D. H. Ballard. Parameter networks: Towards a theory of low level vlsion. In *International Joint Conference on Artificial Intelligence*, pages VI: 1068–1078, 1981.

[4] Adrian Barbu and Song-Chun Zhu. Graph partition by swendsen-wang cuts. In *International Conference on Computer Vision*, page 320, Washington, DC, USA, 2003. IEEE Computer Society.

[5] M. Behrmann, R. Kimchi, and C. Olson, editors. *Perceptual Organization in Vision: Behavioral and Neural Perspectives*. Lawrence Erlbaum Associates, New Jersey, 2003.

[6] Marlene Behrmann and Ruth Kimchi. What does visual agnosia tell us about perceptual organization and its relationship to object perception? *Journal of Experimental Psychology: Human Perception and Performance*, 29(1):19–42, February 2003.

[7] Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste,

and J. Weston, editors, *Large-Scale Kernel Machines*, pages 323–362. MIT Press, Cambridge, Mass., 2007.

[8] J. C. Bezdek, L. O. Hall, and L. P. Clarke. Review of MR image segmentation techniques using pattern recognition. *Medical Physics*, 20(4):1033–1048, 1993.

[9] Bir Bhanu and Sungkee Lee. *Genetic Learning for Adaptive Image Segmentation*. Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[10] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[11] Andrew Blake, Carsten Rother, M. Brown, Patrick Pérez, and Philip H. S. Torr. Interactive image segmentation using an adaptive GMMRF model. In *European Conference on Computer Vision*, pages I: 428–441, 2004.

[12] Andrew Blake and Andrew Zisserman. *Visual reconstruction*. MIT Press, Cambridge, MA, USA, 1987.

[13] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.

[14] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *IEEE Computer Society Workshop on Perceptual Organization in Computer Vision*, page 46, 2004.

[15] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *European Conference on Computer Vision*, page II: 109 ff., 2002.

[16] Eran Borenstein and Jitendra Malik. Shape guided object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 969–976, Washington, DC, USA, 2006. IEEE Computer Society.

[17] Eran Borenstein and Shimon Ullman. Learning to segment. In *European Conference on Computer Vision*, pages 315 – 328, 2004.

[18] A. Bosch, X. Munoz, and J. Freixenet. Segmentation and description of natural outdoor scenes. *Image and Vision Computing*, 25(5):727–740, 2007.

[19] Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient N-D image segmentation. *International Journal of Computer Vision,*, 70(2):109–131, 2006.

[20] Yuri Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *International Conference on Computer Vision*, pages I: 105–112, 2001.

[21] Michael Brady. Computational approaches to image understanding. *ACM Computing Surveys*, 14(1):3–71, 1982.

[22] T. Brox and D. Cremers. On the statistical interpretation of the piecewise smooth Mumford-Shah functional. In *International Conference on Scale Space Methods and Variational Methods in Computer Vision*, pages 203–213, 2007.

[23] Liangliang Cao and Li Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *International Conference on Computer Vision*, pages 1–8, 2007.

[24] Chad Carson, Serge Belongie, Hayit Greenspan, and Jitendra Malik. Blobworld: image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002.

[25] Chad Carson, Megan Thomas, Serge Belongie, Joseph Hellerstein, and Jitendra Malik. Blobworld: a system for region-based image indexing and retrieval. Technical report, University of California at Berkeley, Berkeley, CA, USA, 1999.

[26] T. Chan, J. Shen, and L. Vese. Variational PDE models in image processing. *Notices of the American Mathematical Society*, 50:14–26, Jan. 2003.

[27] Tony Chan and Jianhong Shen. *Image Processing And Analysis: Variational, PDE, Wavelet, And Stochastic Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005.

[28] Junqing Chen, T. N. Pappas, A. Mojsilovic, and B. E. Rogowitz. Adaptive perceptual color-texture image segmentation. *IEEE Transactions on Image Processing*, 14(10): 1524–1536, 2005.

[29] L. D. Cohen. On active contour models and balloons. *Computer Vision, Graphics, and Image Processing*, 53(2):211–218, 1991.

[30] D. Cremers. *Statistical shape knowledge in variational image segmentation*. PhD thesis, Department of Mathematics and Computer Science, University of Mannheim, Germany, 2002.

[31] Daniel Cremers, Mikael Rousson, and Rachid Deriche. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *International Journal of Computer Vision,*, 72(2):195–215, 2007.

[32] Francisco J. Estrada. *Advances in Computational Image Segmentation and Perceptual Grouping*. PhD thesis, Department of Computer Science, University of Toronto, June 2005.

[33] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision,*, 59(2):167–181, 2004.

[34] X. Feng, C. K. I. Williams, and S. N. Felderhof. Combining belief networks and neural networks for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):467–483, 2002.

[35] Mario A. T. Figueiredo, Jose M. N. Leitao Instituto de, and Anil K. Jain. Adaptive B-splines and boundary estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–730, Washington, DC, USA, 1997. IEEE Computer Society.

[36] David Forsyth, Jitendra Malik, Margaret Fleck, and Jean Ponce. Primitives, perceptual organization and object recognition. submitted to Vision Research, Feb. 1997, 1997.

[37] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. Pearson Education Inc., Upper Saddle River, NJ, 2003.

[38] Jordi Freixenet, noz Xavier Mu D. Raba, Joan Martí, and Xavier Cufí. Yet another survey on image segmentation: Region and boundary information integration. In *European Conference on Computer Vision*, pages III: 408–422, 2002.

[39] Brendan J. Frey and Nebojsa Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1392–1416, 2005.

[40] K.S. Fu and J.K. Mui. A survey on image segmentation. *Pattern Recognition*, 13(1):3–16, 1981.

[41] Yoram Gdalyahu. *Stochastic Clustering and its Applications to Computer Vision*. PhD thesis, School of Computer Science and Engineering, Hebrew University in Jerusalem, 1999.

[42] Yoram Gdalyahu, Daphna Weinshall, and Michael Werman. Self-organization in vision: Stochastic clustering for image segmentation, perceptual grouping, and image database organization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1053–1074, 2001.

[43] Feng Ge, Song Wang, and Tiecheng Liu. Image-segmentation evaluation from the perspective of salient object extraction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 1146–1153, 2006.

[44] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[45] Theo Gevers and A. W. M. Smeulders. Content-based image retrieval: An overview. In Gerard Medioni and Sing B. Kang, editors, *Emerging Topics in Computer Vision*, IMSC Press Multimedia Series, chapter 8, pages 333 – 384. Prentice Hall, Upper Saddle River, NJ, 1st edition, 2004.

[46] Greig, D. M., Porteous, B. T., and Seheult, A. H. Exact maximum A Posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(2):271–279, 1989.

[47] Ulf Grenander. A unified approach to pattern analysis. *Advances in Computers*, 10:175–216, 1970.

[48] Ulf Grenander. Tutorial in pattern theory. Technical report, Division of Applied Mathematics, Brown University, 1983.

[49] Ulf Grenander. Advances in pattern theory. *Annals of Statistics*, 17(1):1–30, 1989.

[50] Ulf Grenander. *Elements of pattern theory*. Johns Hopkins University Press, Baltimore, 1996.

[51] Ulf Grenander and Michael I. Miller. *Pattern theory : from representation to inference*. Oxford University Press, Oxford ; New York, 2007.

[52] Peter Hall. Computer vision. Lecture Notes on Computer Vision – version November, 2007.

[53] Xuming He and Richard Zemel. Latent topic random fields: Learning using a taxonomy of labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[54] Xuming He, Richard Zemel, and Miguel Carreira-Perpinan. Multiscale conditional random fields for image labelling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II: 695–702, 2004.

[55] Xuming He, Richard S. Zemel, and Debajyoti Ray. Learning and incorporating top-down cues in image segmentation. In *European Conference on Computer Vision*, pages I: 338–351, 2006.

[56] G. E. Hinton and T. J. Sejnowski. Optimal perceptual inference. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 448 – 453, 1983.

[57] Geoffrey Hinton, Max Welling, Yee-Whye Teh, and Simon Osindero. Learning energy-based models of high-dimensional data. Talk in the Center for Language and Speech Processing Seminar Series, October 2002. Presentation slide.

[58] Thomas Hofmann. Probabilistic latent semantic analysis. In *myUAI*, 1999.

[59] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *myML*, 42(1):177–196, 2001.

[60] D. Hoiem, Carsten Rother, and J. Winn. 3D layoutCRF for multi-view object class recognition and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[61] Berthold K. Horn. *Robot Vision*. MIT Press, Cambridge, Massachusetts, 1986.

[62] A. K. Jain and P.J. Flynn. Image segmentation using clustering. In Kevin Bowyer and Narendra Ahuja, editors, *Advances in Image Understanding: A Festschrift for Azriel Rosenfeld*, pages 65–83. IEEE Computer Society Press, Piscataway, N.J., 1996.

[63] A.K. Jain, Y. Zhong, and M.-P. Dubuisson-Jolly. Deformable template models: A review - algorithms based on Hamilton-Jacobi formulations. *Signal Processing*, 71:109–129, December 1998.

[64] N. Kamath, K.S. Kumar, U.B. Desai, and R. Dugud. Joint segmentation and image interpretation using hidden markov models. In *International Conference on Pattern Recognition*, pages II: 1840–1842, 1998.

[65] Takeo Kanade. Region segmentation: Signal vs. semantics. *Computer Graphics and Image Processing*, pages 279–297, 1980.

[66] A. Kapoor and J. Winn. Located hidden random fields: Learning discriminative parts for object detection. In *European Conference on Computer Vision*, 2006.

[67] A. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contours models. *International Journal of Computer Vision,*, 1:321–331, 1988.

[68] Il Y. Kim and Hyun S. Yang. An integration scheme for image segmentation and labeling based on Markov random field model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):69–73, 1996.

[69] I.Y. Kim and H.S. Yang. An integrated approach for scene understanding based on Markov random field model. *Pattern Recognition*, 28(12):1887–1897, December 1995.

[70] Junmo Kim. *A nonparametric statistical method for image segmentation and shape analysis*. PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2005.

[71] Iasonas Kokkinos and Petros Maragos. An expectation maximization approach to the synergy between image segmentation and object categorization. In *International Conference on Computer Vision*, pages 617–624, 2005.

[72] V. Kolmogorov and Ramin. Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147 – 159, February 2004.

[73] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? In *European Conference on Computer Vision*, pages III: 65–81, 2002.

[74] S. Konishi, A. Yuille, J. Coughlan, and S. Zhu. Statistical edge detection: Learning and evaluating edge cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):57–74, January 2003.

[75] S. M. Konishi and A.L. Yuille. Statistical cues for domain specific image segmentation with performance analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 125–132, 2000.

[76] Scott Konishi, Alan Yuille, and James Coughlan. A statisti-

cal approach to multiscale edge detection. In *International Workshop on Generative-Model Based Vision*, 2002.

[77] K. Sunil Kumar and U. B. Desai. Joint Segmentation and Image Interpretation. In *International Conference on Image Processing*, September 1996.

[78] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Learning layered pictorial structures from video. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pages 158–163, 2004.

[79] M. P. Kumar, P. H. S. Torr, and A. Zisserman. An object category specific MRF for segmentation. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 596–616. Springer, 2006.

[80] M. Pawan Kumar, Philip H. S. Torr, and A. Zisserman. OBJ CUT. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 18–25, 2005.

[81] M.P. Kumar, P.H.S. Torr, and A. Zisserman. Extending pictorial structures for object recognition. In *British Machine Vision Conference*, 2004.

[82] Sanjiv Kumar and Martial Hebert. Discriminative random fields. *International Journal of Computer Vision,*, 68(2): 179–201, 2006.

[83] Yann. LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. Tutorial on energy-based learning. In Gükhan H. Bakir, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar, and S. V. N.

Vishwanathan, editors, *Predicting Structured Data*, Neural Information Processing, pages 191–246. The MIT Press, 2007.

[84] Yann LeCun, Sumit Chopra, Marc'Aurelio Ranzato, and Fu-Jie Huang. Energy-based models in document recognition and computer vision. In *International Conference on Document Analysis and Recognition*, pages 337–341, 2007. (keynote address).

[85] Yann LeCun and Fu Jie Huang. Loss functions for discriminative training of energy-based models. In *International Workshop on Artificial Intelligence and Statistics*, 2005.

[86] Sébastien Lefévre, Jérôme Holler, and Nicole Vincent. A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real-Time Imaging*, 9(1):73–98, 2003.

[87] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV Workshop on Statistical Learning in Computer Vision*, pages 17–32, 2004.

[88] B. Leibe, K. Mikolajczyk, and B. Schiele. Segmentation based multi-cue integration for object detection. In *British Machine Vision Conference*, page III: 1169, 2006.

[89] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *British Machine Vision Conference*, pages 759–768, Norwich, UK, Sept. 2003.

[90] B. Leibe and B. Schiele. Scale invariant object categorization using a scale-adaptive mean-shift search. In *Annual*

Symposium of the German Association for Pattern Recognition, pages 145–153, 2004.

[91] B. Leibe and B. Schiele. Interleaving object categorization and segmentation. In *Cognitive Vision Systems – Sampling the Spectrum of Approaches*, Lecture Notes on Computer Science, 3948, pages 145–161. Springer, 2006.

[92] Bastian Leibe. *Interleaved Object Categorization and Segmentation*. PhD thesis, ETH Zurich, October 2004.

[93] Erich Leung and John Tsotsos. Probabilistic representation and inference in natural scene segmentation. Technical report, York University, Toronto, Canada, 2009.

[94] Anat Levin and Yair Weiss. Learning to combine bottom-up and top-down segmentation. In *European Conference on Computer Vision*, pages 581–594, 2006.

[95] Yin Li, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Lazy snapping. In *ACM SIGGRAPH: International Conference on Computer Graphics and Interactive Techniques*, pages 303–308, 2004.

[96] Pan Lin, Chong-Xun Zheng, Yong Yang, and Jian-Wen Gu. Statistical model based on level set method for image segmentation. *International Conference on Computer and Information Technology*, pages 143–148, 2004.

[97] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, Jan 2007.

[98] D. Lu and Q. Weng. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5): 823–870, 2007.

[99] Jitendra Malik, Serge Belongie, Jianbo Shi, and Thomas K. Leung. Textons, contours and regions: Cue integration in image segmentation. In *International Conference on Computer Vision*, pages II: 918–925, 1999.

[100] D. Marr. Early processing of visual information. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 275(942):483–519, Oct 1976.

[101] D. Marr, S. Lal, and H. B. Barlow. Visual information processing: The structure and creation of visual representations. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 290(1038):199–218, July 1980.

[102] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982.

[103] Joan Marti, Jordi Freixenet, Joan Batlle, and Alicia Casals. A new approach to outdoor scene description based on learning and top-down segmentation. *Image and Vision Computing*, 19(14):1041–1055, December 2001.

[104] D. Martin. *An Empirical Approach to Grouping and Segmentation*. PhD thesis, University of California, Berkeley, December 2002.

[105] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms. In *IEEE Computer Society Workshop on Perceptual Organization in Computer Vision*, 2001.

[106] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images its application to evaluating segmentation algorithms and measuring ecological statistics. In *International Conference on Computer Vision*, pages 416–425, 2001.

[107] David R. Martin, Charless C. Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004.

[108] R. Miikkulainen, J. A. Bednar, Y. Choe, and J. Sirosh. *Computational maps in the visual cortex.* Springer, New York, NY, 2005.

[109] J.W. Modestino and J. Zhang. A Markov random field model-based approach to image interpretation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(6):606–615, 1992.

[110] A. Mojsilovic and B. Rogowitz. Capturing image semantics with low-level descriptors. In *International Conference on Image Processing*, pages I: 18–21, 2001.

[111] E. N. Mortensen. *Simultaneous Multi-Frame Subpixel Boundary Definition using Toboggan-Based Intelligent Scissors for Image and Movie Editing.* PhD thesis, Department

of Computer Science, Brigham Young University, December 2000.

[112] E. N. Mortensen and W. A. Barrett. Intelligent scissors for image composition. In *ACM SIGGRAPH: International Conference on Computer Graphics and Interactive Techniques*, pages 191–198, 1995.

[113] E. N. Mortensen and W. A. Barrett. Interactive segmentation with intelligent scissors. *Graphical Models and Image Processing*, 60(5):349–384, September 1998.

[114] E.N. Mortensen, L.J. Reese, and W.A. Barrett. Intelligent selection tools. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II: 776–777, 2000.

[115] Eric N. Mortensen and William A. Barrett. Toboggan-based intelligent scissors with a four-parameter edge model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II: 2452–2458, 1999.

[116] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Communications On Pure & Applied Mathematics*, 42:577–685, 1989.

[117] Makoto Nagao. Control strategies in pattern analysis. *Pattern Recognition*, 17(1):45–56, 1984.

[118] R. Nevatia. *Machine Perception*. Prentice-Hall, Englewood Cliffs, NJ, 1982.

[119] S. L. Ngohayon, J. Kawhara, and T Toshima. Is visual image segmentation affected by higher-level object representation? In *Joint Conference of International Conference*

of Cognitive Scince/Japanese Cognitive Science Society Annual Meeting, Japanese Cognitive Science Society, 1999.

[120] M. Nitzberg, D. Mumford, and T. Shiota. *Filtering, Segmentation, and Depth*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1993.

[121] T. N. Pappas, J. Chen, and D. Depalov. Perceptually based techniques for image segmentation and semantic classification. *IEEE Communications Magazine*, 45:44–51, January 2007.

[122] Nikos Paragios and Rachid Deriche. Coupled geodesic active regions for image segmentation: A level set approach. In *European Conference on Computer Vision*, pages 224–240, 2000.

[123] Nikos Paragios and Rachid Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *International Journal of Computer Vision,*, 46(3):223–247, 2002.

[124] Patrick Perez, Andrew Blake, and Michel Gangnet. JetStream: Probabilistic contour extraction with particles. In *International Conference on Computer Vision*, pages II: 524–531, 2001.

[125] M. A. Peterson. Organization, segregation and object recognition. *Intellectica*, 28(3/4):37 – 51, 1999.

[126] M. A. Peterson and B. S. Gibson. Must figure-ground organization precede object recognition? an assumption in peril. *Psychological Science*, 5(5):253–259, 1994.

[127] M. A. Peterson and B. S. Gibson. Object recognition contributions to figure-ground organization: Operations on outlines and subjective contours. *Perception & Psychophysics*, 56(5):551–564, 1994.

[128] N. Pinto, D.D. Cox, and J. J. DiCarlo. Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4(1):e27, 2007.

[129] Axel Pinz. Object categorization. *Foundations and Trends in Computer Graphics and Vision*, 1(4), 2006.

[130] M. Ranzato, Y. Boureau, S. Chopra, and Y. LeCun. A unified energy-based framework for unsupervised learning. In *International Workshop on Artificial Intelligence and Statistics*, 2007.

[131] X. Ren, C. Fowlkes, and J. Malik. Figure/ground assignment in natural images. In *European Conference on Computer Vision*, 2006.

[132] Xiaofeng Ren, Charless C. Fowlkes, and Jitendra Malik. Cue integration in figure/ground labeling. In *Advances in Neural Information Processing Systems*, 2005.

[133] Xiaofeng Ren, Charless C. Fowlkes, and Jitendra Malik. Mid-level cues improve boundary detection. Technical Report UCB//CSD-05-1382, UC Berkeley, 2005.

[134] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *International Conference on Computer Vision*, pages I: 10–17, 2003.

[135] Jordan Reynolds and Kevin Murphy. Figure-ground segmentation using a hierarchical conditional random field. In *Canadian Conference on Computer and Robot Vision*, 2007.

[136] E.M. Riseman and M.A. Arbib. Computational techniques in the visual segmentation of static scenes. *Computer Graphics and Image Processing*, 6(3):221–276, June 1977.

[137] Edmund T. Rolls. *Memory, attention, and decision-making: A unifying computational neuroscience approach.* Oxford University Press, Oxford, 2008.

[138] Azriel Rosenfeld and Avinash C. Kak. *Digital picture processing*, volume 1 and 2. Academic Press, New York, 2nd edition, 1982.

[139] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. GrabCut: interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004.

[140] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1605–1614, Washington, DC, USA, 2006. IEEE Computer Society.

[141] Jayant Shah. Properties of energy-minimizing segmentations. *SIAM Journal on Control and Optimization*, 30(1): 99–111, 1992.

[142] Linda G. Shapiro and George Stockman. *Computer Vision.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[143] E. Sharon, A. Brandt, and R. Basri. Fast multiscale image segmentation. In *International Conference on Computer Vision*, pages 70–77, 1999.

[144] E. Sharon, A. Brandt, and R. Basri. Segmentation and boundary detection using multiscale intensity measurements. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 469–476, 2001.

[145] E. Sharon, M. Galun, D Sharon, R Basri, and A. Brandt. Hierarchy and adaptivity in segmenting visual scenes. *Nature*, 442:810–813, August 2006.

[146] J. Shi and J. Malik. Normalized cuts and image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 731–737, 1997.

[147] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.

[148] Jianbo Shi and Jitendra Malik. Normalized cut and image segmentation. Technical Report UCB/CSD-97-940, EECS Department, University of California, Berkeley, 1997.

[149] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision*, 2006.

[150] Wesley Snyder and Hairong Qi. *Machine Vision*. Cambridge University Press, New York, NY, USA, 2003.

[151] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image Processing, Analysis, and Machine Vision*. Thomson-Engineering, 2007.

[152] J.M. Tenenbaum and H.G. Barrow. Experiments in interpretation guided segmentation. *Artificial Intelligence*, 8(3): 241–274, June 1977.

[153] Sinisa Todorovic and Narendra Ahuja. Extracting subimages of an unknown category from a set of images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 927–934, Washington, DC, USA, 2006. IEEE Computer Society.

[154] Yaakov Tsaig. Automatic segmentation of moving objects in video sequences. Master's thesis, Tel-Aviv University, 2001.

[155] J.K. Tsotsos. Image understanding. In S. Shapiro, editor, *The Encyclopedia of Artificial Intelligence*, pages 641 – 663. John Wiley and Sons, second edition, 1992.

[156] Zhouwen Tu, Xiangrong Chen, Alan L. Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. Technical report, Department of Statistics, UCLA. Department of Statistics, January 2005.

[157] Zhuowen Tu, Xiangrong Chen, Alan L. Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision,*, 63(2):113–140, 2005.

[158] Zhuowen Tu and Song-Chun Zhu. Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, 24(5):657–673, May 2002.

[159] Zhuowen Tu and Song-Chun Zhu. Parsing images into regions, curves, and curve groups. *International Journal of Computer Vision,*, 69(2):223–249, 2006.

[160] S. P. Vecera and R. C. O'Reilly. Figure-ground organization and object recognition processes: An interactive account. *Journal of Experimental Psychology: Human Perception and Performance*, 24(2):441–462, 1998.

[161] O. Veksler. *Efficient graph-based energy minimization methods in computer vision.* PhD thesis, Department of Computer Science, Cornell University, 1999.

[162] B. Verbeek, J.and Triggs. Region classification with Markov field aspect models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 – 8, 2007.

[163] Jakob Verbeek and Bill Triggs. Scene segmentation with conditional random fields learned from partially labeled images. In *Advances in Neural Information Processing Systems*, volume 20, pages 1553–1560, 2007.

[164] J. Winn and N. Joijic. LOCUS: Learning object classes with unsupervised segmentation. In *International Conference on Computer Vision*, pages I: 756–763, 2005.

[165] John Winn and Jamie Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 37–44, 2006.

[166] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113, 1993.

[167] C. Zhang and T. Chen. From low level features to high level semantics. In Borko Furht and Oge Marques, editors, *The Handbook of Video Database Design and Applications*, pages 613 – 624. CRC Press, Routledge, Sept 2003.

[168] D. S. Zhang and G. Lu. Segmentation of moving objects in image sequence: A review. *Circuits, Systems and Signal Processing*, 20(2):143–183, 2001. Special Issue on Multimedia Communication Services.

[169] Y. J. Zhang. An overview of image and video segmentation in the last 40 years. In *Advances in Image and Video Segmentation*. IRM Press, Hershey, PA., 2006.

[170] Song-Chun Zhu. Statistical modeling and conceptualization of visual patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):691– 712, June 2003.

[171] Song Chun Zhu and Alan Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996.

[172] Thomas Zöller and Joachim M. Buhmann. Robust image segmentation using resampling and shape constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1147–1164, July 2007.