# A Syntax for Image Understanding

Narendra Ahuja
Department of Electrical and Computer Engineering, Coordinated Science Laboratory, and
Beckman Institute
University of Illinois at Urbana-Champaign

We consider one of the most basic questions in computer vision, that of finding a low-level image representation that could be used to seed diverse, subsequent computations of image understanding. Can we define a relatively general purpose image representation which would serve as the syntax for diverse needs of image understanding? What makes good image syntax? How do we evaluate it? We pose a series of such questions and evolve a set of answers to them, which in turn help evolve an image representation. For concreteness, we first perform this exercise in the specific context of the following problem.

**Problem:** We wish to discover *a priori* unknown themes that may characterize a given set of images. The images may be selected randomly, in which case any discovered theme will capture accidental commonalities among the properties of the images selected. Alternatively, the images may be chosen strategically, known to possess some common attribute, e.g., they may all depict indoor scenes. In this latter case, any extracted theme will provide an image-space definition of the scene type. We define a theme in terms of the objects present in the set. If objects from certain categories occur frequently in the set, we say that the categories constitute the theme. No specific categories are specified in advance; indeed, they are not even known *a priori*. Whether, how many, or where instances of any categories appear in a specific image is also not known.

**Objects and Categories:** To solve the aforementioned problem, we develop answers to the following basic questions. What are the definitions of objects present in images? What is an object category? What properties should be used to define a good category representation? If, and to what extent, is human supervision necessary to communicate the nature of categories to a computer vision system?

Here we restrict our attention to 2D objects. We argue that since many objects of interest in the real world are spatially localized, collections of appropriately selected (2D) subimages can be used to define the regions of objects' projections in images, and therefore, image regions serve as good primitives for defining objects. The subimages can be characterized in terms of low-level image properties.

**Syntax:** We argue that all subimages of relevance here, i.e., those suited to form real world objects, are those contained in a multiscale low-level segmentation of the image. The segmentation algorithm should be able to extract all perceptually salient regions present, irrespective of their geometric and photometric complexity and scales. The properties of these regions constitute the low level properties needed to form category definitions. A representation of all the regions and their necessary properties define our syntax for image understanding. This is because this representation serves as a sufficient parse of the image that, because of the natural perceptual-region vs. object relationships, meets the needs of subsequent object-oriented image understanding computations.

The specific syntax we present is called Connected Segmentation Tree (CST), which is defined in terms of intra- as well as inter-region properties of image segments. It consists of three parts:

(a) The nodes in the tree correspond to the extracted multiscale image segments, and capture their geometric and photometric properties.

(b) The parent child relationships capture recursive embedding (containment) of regions.

(c) Additional links among tree nodes specify regions' neighbors in all directions, thus capturing spatial layout of regions. This part extends the segmentation tree of parts (a) and (b) into a graph, by adding the neighbor connections.

We summarize the derivation of CST from images, and its invariance to changes in imaging conditions (e.g., lighting, scale, orientation).

**Examples of Object Level Image Understanding:** We define an object category as consisting of subimages that have similar photometric, geometric and topological properties.

We evaluate the utility of the CST syntax in inferring semantics, as would be expected from any syntax. We do so by using the CST to solve the following object-oriented image understanding problems:

(1) *Discovering* whether any categories occur in the image set. This amounts to the subtree isomorphism problem.

(2) *Building* a compact model that captures the intrinsic nature of the categories. This can be achieved by capturing a distribution of matching subtree instances in a single connected tree of underlying probability distributions.

(3) *Learning* the relationships among the different categories, thus building a taxonomy of all discovered categories. This requires discriminating between sets of trees corresponding to within-category and cross-category instances of objects.

(4) *Recognizing* all occurrences of all categories in previously unseen images, using the learned taxonomy. This is analogous to problem (1), now applied to matching a model tree obtained in (2) with the segmentation trees of the unseen images.

(5) *Segmenting* each occurrence. Along with recognition, a matched subtree also segments the recognized object, by definition of the segmentation tree.

(6) *Explaining* the reasons for recognition. The labels of the matched nodes and the parent-child and neighbor links in the model CST provide an explanation of why a category has been detected, in terms of where and which constituent subcategories within the model CST are detected.

We summarize our solutions to (1–6), all of which are almost completely unsupervised.

**Generality of the CST Syntax:** Syntax should be useful for multiple types of semantic inferences. The general nature of the problems (1-6) helps extend their solutions to detecting themes of other kinds. As two examples of evaluating such generality, we summarize our solutions to the following two related problems.

(1) *Texture Element Extraction:* Texture elements, or texels, define a theme within a texture image, because they correspond to stochastic recurrences of the same object within the image. We evaluate the performance of CST in terms of the quality of the detected texels in real-world textures.

(2) *Texture Segmentation:* In an image consisting of multiple textures, we identify the multiple texture “categories” present, each defined by different type of texel. As a result, we can segment the image into its different parts occupied, by the different estimated texture types.