# Complexity: The Fundamental Problem of Vision

- Complexity is the fundamental problem of vision. How can the visual system manage to represent all possible objects and recognize them quickly from an input image (how to learn is also fundamental but we will ignore this until later).
- We will study this from the perspective of hierarchical compositional models. These models are less effective than deep networks but they are easier to analyze. They differ by being semantic and by representing objects in terms of parts which are recursively composed of more elementary parts. The most elementary parts can be oriented edge bars.
- For example, the letter T can be represented in terms of a vertical bar and a horizontal bars satisfying spatial relationships (i.e. the vertical bar is below the horizontal bar and intersects it near the center). Similarly the letter L is also represented in terms of vertical and horizontal bars, but obeying different spatial relations.
- Intuitively, the model consists of a hierarchy of dictionaries (analagous to the filterbanks in deep network). Dictionary elements at one level are composed of compositions of dictionary elements at the previous layer. The number of dictionary elements at each layer is variable (unlike deep networks) and will be learnt.

# Hierarchical Dictionaries: Executive Summary

- The dictionary elements are encoded as part detectors defined over different lattices (ie.e. we have part detectors at every position on each lattice). The lattices become increasingly coarse as we ascend the hierarchy. A part detector at one level can be activated by several different subparts being detected at lower levels. This enables the part detector to only coarsely represent the position of the part (i.e. on the coarse lattice) because the detailed locations of the subparts are available at the previous level.
- This is called the executive summary principle. The higher levels of the model represent the coarse description of the objects – i.e. there is a horse in a field – while the lower levels represent the details of the positions of the parts of the horse, the subparts, and ultimately the boundary of the horse.
- This is analogous to a hierarchical organization like an army or a corporation. The general of the army has executive summary information only and does not know the details. But the general knows to get details by consulting the colonels, who can consult the majors, and so on, until reaching the private soldiers. This enables information to be stored throughout the hierarchy in an efficient manner. For vision, this means we can store and access a very large numbers of objects and parts/subparts in a hierarchical architecture.

# Inference: Propagating Hypotheses

- Inference can be performed rapidly in a bottom-up and top-down loop. The bottom-up process propagates hypotheses up the hierarchy by computing the evidence for parts at different locations. This evidence is computer recursively from the evidence for the subparts and their spatial relations. At each position, the algorithm selects the maximum response from all it possible subpart configurations. This select the best sub-part configuration and "binds" it to the part. The only sub-part configurations are suppressed.
- More formally, the algorithm proceeds by dynamic programming with a forward (bottom-up stage) and a backwards stage. This is analogous to the inside-outside algorithm used for parsing sentences in Natural Language Processing (NLP).
- This enables the optimal estimate to be performed in a single bottom-up and top-down pass. The bottom-up pass gives the estimates of the high-level parts (and the object identify) very quickly. The top-down pass eliminates the low-level hypotheses which are inconsistent with the top-level hypotheses.

## Learning the Dictionaries

- The dictionaries can be learnt in an unsupervised manner (at least in principle, it is very hard in practice). This is done by recursive clustering. This starts with the elementary dictionary elements which are pre-specified, e.g, edge detectors at different orientations.
- To start this process, we run the lowest-level part detectors over a set of images. Then we search for clusters of spatial configurations of these lowest-level parts which occur frequently. These are selected to be the dictionary elements for the next layer.
- For example, suppose we have a set of images made from random samples of vertical and horizontal bars with a few T's and L's. The clustering algorithm will detect the T's and L's as "suspicious coincidences" (unlikely to occur by change as randomly placed vertical and horizontal bars. The lowest level dictionary consists of horizontal and vertical bars. The next level consists of T's and L's.
- This process was run on 120 silhouettes of objects. It produced a series of dictionaries where the top-dictionary corresponded to the objects. This was a hierarchical model where the depth of the hierarchy was determined by the data (the learning algorithm stopped when it failed to find any more suspicious coincidences).

# Capacity of the Architecture

- This gives a hierarchical architecture which can represent and rapidly infer objects, and object parts, in terms of hierarchical dictionaries. The capacity of the architecture depends on *part sharing* and *executive summary*. The more we can share parts between different objects the smaller the size of the dictionaries (and hence the more efficient). The more we can use executive summary (by representing parts more sparsely in space) then the more efficient the architecture.
- Theoretical analysis shows that this architecture can be extremely efficient.
- What is the take-away message? Compositional models have many conceptual advantages but remain mush less effective than deep networks (they are learnt by more complex algorithms) This type of detailed analysis can be done on hierarchical compositional models because of their semantic structure (explicit representations of parts) but, at present, this explcitness/explainability decreases performance compared to deep networks (the so-called explainability gap).