Learning Hierarchical Compositional Representations of Object Structure

Sanja Fidler Marko Boben Aleš Leonardis
University of Ljubljana, Slovenia

1 Introduction

Visual categorization of objects has captured the attention of the vision community for decades [10]. The increased popularity of the problem witnessed in the recent years and the advent of powerful computer hardware have led to a seeming success of categorization approaches on the standard datasets such as Caltech 101 [15]. However, the high discrepancy between the accuracy of object classification and detection/segmentation [14] suggests that the problem still poses a significant and open challenge. The recent preoccupation with tuning the approaches to specific datasets might have precluded the attention from the most crucial issue: the representation [41].

This paper will focus on what we believe are two central representational design principles, namely a hierarchical organization of categorical representations, more specifically, the principle of *hierarchical compositionality*, and *statistical, bottom-up learning*.

Given images of complex scenes, objects must be inferred from the pixel information through some recognition process. This requires an efficient and robust matching of the internal object representation against the representation produced from the scene. Despite the seemingly effortless performance of human perception, the diversity and the shear number of visual object classes appearing in various scales, 3D positions and articulations, which additionally interact with each other (occlusion, clutter, etc.), have placed a great obstacle to the task. In fact, it has been shown by Tsotsos in 1990 [53] that the unbounded visual search is NP complete and thus approximate, hierarchical solutions might be the most promising/plausible way to tackle the problem. This line of architecture is also consistent with the findings on biological systems [44, 9]. A number of authors have further emphasized these computational considerations [13, 23, 2, 47, 26], suggesting that matching should be performed at multiple hierarchical stages, in order to gradually and coherently limit the otherwise computationally prohibitive search space [13, 53, 8, 3, 33, 2, 23, 17, 47, 19]. While hierarchies presented a natural way to represent objects in the early vision works [13, 24, 32, 11], surprisingly, they have not become an integral part of the modern vision approaches.

Hierarchical representations can derive and organize the features at multiple levels that build on top of each other by exploiting the shareability of features among more complex compositions or objects themselves [13, 7, 21, 2,

30, 25, 51, 38, 55]. Sharing features, on the one hand, means sharing common computations, which brings about the much desired computational efficiency. On the other hand, reusing the commonalities between objects can put their representations in relation, thus possibly leading to high generalization capabilities [13, 30, 51]. A number of hierarchical recognition systems have been proposed and confirmed the success of such representations in object categorization tasks [22, 3, 21, 48, 45, 1, 47, 56, 34, 36, 51, 46, 42, 38, 50].

It must be emphasized, however, that hierarchical representations do not necessarily imply computational efficiency and representational plausibility. In this paper, we will argue for a special form of a multilayered architecture — a *compositional hierarchy*. The nodes in such a hierarchy are formed as compositions that, recursively, model loose spatial relationships between their constituent components. The abundant computational arguments accumulated throughout the history of computational vision speak in favor of its efficiency, robustness to clutter, and flexibility to capture structural variability. While the classical neural networks have been commonly thought to be a faithful model of the hierarchical processing in the brain, interestingly, ideas of compositional units have also started to emerge in the neuroscience community [4, 9, 54].

There have been a number of attempts at compositional categorical representations [13, 11, 23, 8, 3, 2], however, the lack of automation might have been a major contributing factor that prevented better realizations of these ideas. It is, in fact, the absence of *unsupervised, bottom-up learning* principles that also seems to be the source of criticism by the neuroscience community targeted at today's computational models of vision [39]. By statistically learning the priors to bindings of local features, the representation emerging in this way would be well adjusted to the regularities present in images and could thus reliably, robustly and quickly form hierarchical groupings facilitating the final recognition of objects.

Based upon the computational considerations of compositional hierarchies and benefits of the bottom-up, unsupervised learning, we will summarize our recent approach to representing object categories within a *learnable, hierarchical compositional framework*. The developed *bottom-up, statistical approach* makes use of simple atomic features, i.e. oriented edges, to gradually learn more complex contour compositions that model loose spatial relations between the constituent features. The learned hierarchical vocabulary of features, termed *parts*, is organized in accordance with the principle of efficient indexing. This ensures that local retrieval of compositional models during the online object recognition stage runs in a roughly constant time — despite an exponential increase in the number of vocabulary features along the hierarchical layers. An off-line grouping stage of part labels brings additional flexibility into the learned representation. The learned contour compositions can be further combined into categorical nodes with minimal human supervision, whereby the hierarchical sharing of features and the efficient indexability constraints could present an important step towards scalable representations of object categories.

The remainder of this chapter is organized as follows. In Section 2 we argue for the design principles we believe are important to build plausible representation of object structure. Section 3 reviews the work most related to ours. Section 4 is more technical and summarizes our recently developed approach to learning hierarchical compositional representations of object structure. The experimental results are presented in Section 5. The chapter concludes with a

summary and discussion in Section 6.

2 Design principles for representing object structure

A representation must drive the recognition process from the pixel level through more and more complex interpretations towards object categories themselves. It is thus critical to devote our attention to design principles that would accommodate for scalability and generalizability of the representation and robustness as well as efficiency of the subsequent recognition. This Section brings forward two issues we believe are crucial for forming plausible representations of object structure.

We will argue for *hierarchical compositionality* as the line of representational architecture and the principle of *unsupervised, bottom-up learning* that statistically extracts the multiple layers of representation from images making it adjustable to the regularities present in the otherwise highly variable structure of objects.

2.1 Hierarchical compositionality

Compositionality refers to a property of hierarchical representational systems that define their internal nodes in terms of simpler constituent components according to a set of production rules [7]. The rules of composition usually take the form of the Gestalt laws of grouping [37, 16] or similar forms of predefined bindings [13, 3, 47, 50, 59] that in some form or another incorporate spatial relations into the compositional features. Computational benefits of compositionality in terms of storage, processing demands, robustness to clutter and the exponential expressive power have long been emphasized in the computer vision literature [6, 3, 23, 58, 59, 19, 8]. We substantiate these issue below.

Storage demands. In the current state-of-the-art flat representations millions of distinctive image patches (with dimensions ranging around 25×25 pixels) or local descriptors such as SIFT or HoG must be stored to produce good recognition results. In the classical hierarchies such as neural networks the number of necessary features to warrant competitive performance is grantedly significantly lower (the number ranges from $10 - 20$ in the lowest layer and increases to the order of a few thousand in the top-most layer), however, each hierarchical unit still must encode weights to *all* feature types from a layer below covering a certain spatial neighborhood. Conversely, as the complexity and size or representation also grows with the number of layers in compositional hierarchies, each higher-level composition encodes only pointers to a small number of its constituent parts and a modest amount of additional information binding the parts spatially. Furthermore, since all the higher-level compositions are constructed from a smaller common vocabulary from a layer below, it is easier to compare and generalize between them. Consequently, extending the hierarchical library to novel compositions can operate in a more controlled manner leading to more compact and parsimonious hierarchical vocabularies.

Processing complexity. Since each hierarchical unit is shared among many more complex higher layer compositions, most of the computations performed during an online recognition stage are inherently common and can thus

be only performed once. Such sharing of computations greatly reduces the computational cost of matching with respect to searching for each complex interpretation in isolation. Moreover, as processing of images is done by sequential (hierarchical) testing of compositional hypotheses, recognition towards the final categorical nodes proceeds in a more controlled and fast manner by pruning the object hypothesis space along the hierarchical path.

The important advantage of discrete representations such as the compositional architectures is also the possibility of implementing the *indexing and matching scheme*. Since each internal node of the hierarchical vocabulary participates in only a smaller subset of all compositions from the incident layer above, only this specific subset needs in fact be matched against the local image neighborhood during the online recognition stage. Consequently, retrieval of permitted composite models can be performed in constant time, the process termed *indexing*, while the verification of the retrieved candidates runs in sublinear time with respect to the size of the hierarchical library. This procedure will be described in more detail in Subsection 4.2.

Robustness to clutter, repeatability of detection. Each hierarchical node makes inference over a certain size of a local neighborhood, usually referred to as its *receptive field*. The level of hierarchy brings about larger and larger portions of an image that the nodes "cover" and which are likely to contain many structures pertaining to different objects in the scene or rarer structures that the hierarchical units are not essentially tuned to. The classical neural networks that define the units as some non-linear function of an integrative weighted sum over its entire receptive field both spatially as well as in all constituent feature types (schematically depicted in Figure 1) are inherently prone to error since the signal coming from multiple objects is essentially mixed. This can be alleviated by enforcing sparsity on the feature weights to enable a focus on only particular substructures of receptive fields. In turn, compositions are inherently sparse — they are designed to respond to only small spatial subsets of their receptive fields in which the presence of only a few feature types is accounted for (depicted in Figure 1). This ensures that clutter has little effect on the activity within the hierarchical recognition process and additionally permits faster processing over the traditional neural networks approaches.

Expressive power. Even a small number of feature types defining the outset of the hierarchy can construe a large number of possible combinations, which becomes even more pronounced with the level of hierarchy. Importantly, as the vocabulary is expected to grow with exponential tendency as new layers (compositions of compositions that essentially should converge to objects themselves) are added and the complexity as well as distinctiveness of the representation increase, the principle of indexing ensures tractability of the recognition process.

Feedforward and feedback. There has been a long-standing debate about what can or cannot be achieved in a strictly feedforward manner in vision in general and in hierarchical categorization approaches in particular [49, 28, 31, 52]. There is neurophysiological evidence proving good categorization performance in the first feedforward pass by humans [49, 57], while many authors emphasize the importance of both, feedforward and feedback and the iterative process between the two [28, 31, 55, 52]. The ability to traverse back from the final recognition nodes inferred from the scene back to the original pixels that produced the high-level decision is important for segmentation as well as looping between bottom-up and top-down inference on ambiguous visual input. This

kind of reciprocal inference presents an impediment for the neural network approaches [27] and their closely related architectures [43] since the firing response of a hierarchical unit is too reductive. The information from a cube-like receptive field over lower-layer feature responses is conveyed in only one value — the weighted sum. This makes it difficult to determine and trace back what has in fact caused the response (depicted in Figure 1) while also making inference less controlled and reliable. Conversely, in compositional architectures the representation inferred from the visual scene is essentially a graph in which each node has only a small number of incident descendants. Such a representation inherently allows for iterative loops between the data (image) and high-level inferences, whereby the segmentation of objects is simply an inverse process of recognition.

A part of the biological evidence could potentially support such a line of compositional architecture [40, 54, 29, 4, 9]. Additionally, attempts have been made to map the mathematical theory of compositionality onto the neuronal structure of the visual cortex [7].

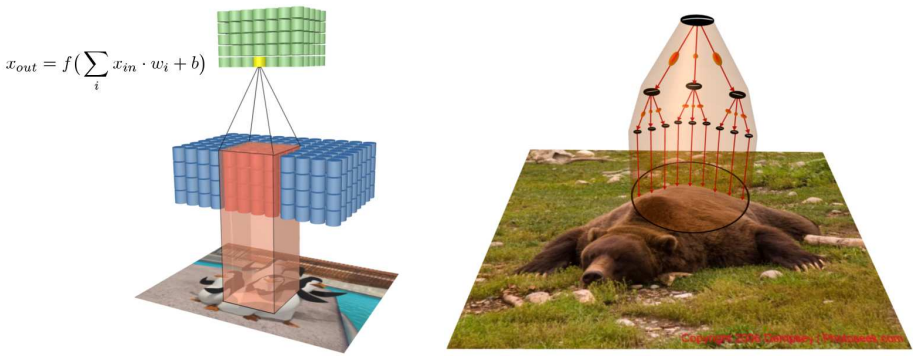

Figure 1: Left: Neural networks. Right: Hierarchical compositionality.

2.2 Statistical, bottom-up learning

The appealing properties of compositional hierarchies and their advantages over the related hierarchical architectures might prove them a suitable form of representing visual information. However, while learning presents an integral part of the neural networks approaches, most compositional approaches have been hindered by the use of predetermined sets of features or grouping rules. Here, we argue for the importance of learning, specifically, we emphasize the critical role of unsupervised, bottom-up learning.

Bottom-up learning. There seems to be a consensus that the higher-level concepts such as selectivity to object categories are learned since, evidently, a genetic predisposition towards e.g. mobile phones and similar ever-evolving technological gadgets would seem far-fetched. Interestingly, there are opposing views on whether the tunings in the early cortical areas are learned or hard-wired by evolution. The diverse physiology underlying different brain areas suggests specific functionalities and computations performed. This striking systematicity surely is a result of evolution and it undoubtedly guides and controls what the cells can or cannot become tuned to. However, it is highly improbable that all

the low-level sensitivities are instilled genetically — the brain must, after all, adjust its perceptual functioning with respect to its sensory receptors and the input it receives.

Computationally speaking, the categorical representations are built upon a set of features that must at some point operate on the image data. The design of these features (for example corners, T- and L-junctions, etc) should not rely on our intuition but rather be learned from the data in order to conform well to the local structures of images. The features/models in the lowest level of the hierarchy should thus be brought down close to the images by performing simple operations with little semantic value. The subsequent learning should then be designed in order to statistically build more complex and semantic models in composition.

Once the *visual building blocks* are learned, learning of objects becomes tractable since only a small number of descriptive structural features are needed to explain them away. Thus, categorical learning can proceed mainly in the higher hierarchical layers and can thus operate fast and with no or minimal human supervision.

Unsupervised learning. Features and their higher level combinations should be learned in an unsupervised manner (at least in the first stages of the hierarchy) in order to avoid hand-labeling of massive image data as well as to capture the regularities within the visual data as effectively and compactly as possible [5, 44, 12, 25, 17, 19]. Moreover, there are strong implications that the human visual system is driven by these principles as well [20].

By learning the compositional binding priors the representation becomes adjustable to the structural variability of objects. Consequently, it enables a computationally feasible recognition process where the majority of the exponential number of possible compositional groupings are made unimportant (i.e. unrepeatable) by the statistics of natural images.

Incremental learning. Desirably, the hierarchical vocabulary should be extended incrementally as new images/objects are seen by the system. Performed in this way, we avoid batch processing of masses of images (which likely might not even be possible), while on the other hand, we ensure the representation is open to continuous adaptation of the visual environment.

The issue of incrementality in hierarchical architectures is not completely apparent. If features are changed, removed or added at any layer exclusive of the top-most one, all the features on the layers above must be adjusted accordingly. This problem is particularly evident in the neural network type of hierarchies where adding one feature results in the inefficient restructuring of the weights of the complete representation. In compositional hierarchies this problem concerns only a small subset of higher-level features that compositionally emerge from the point of change. Furthermore, by learning the representation sequentially, i.e. optimally adjusting layer after layer to the regularities present in the natural signals, we guarantee that very little encoded information (if something at all) will need to be re-adapted in the deepest layers of the hierarchy as new data is encountered.

3 Related work

The current state-of-the-art categorization methods predominantly build their representations on image patches [34, 56] or other highly discriminative features such as the SIFT [48]. Since the probability of occurrence of such features is very small, masses of them need to be extracted to represent objects reasonably well. This results in computationally highly inefficient recognition, which demands matching of a large number of image features to enormous amounts of prototypical ones. This drawback has been alleviated within the most recent methods that employ hierarchical clustering in a high dimensional feature space, yet the resulting representations still demand at least a linear search through the library of stored objects/categories [34, 48].

To overcome the curse of large-scale recognition, some authors emphasized the need for indexable hierarchical representations [8, 3, 2]. A hierarchy of parts composed of parts that could limit the visual search by means of indexing matching in each individual layer would enable an efficient way to store and retrieve information.

However, a majority of hierarchical methods perform matching of *all* prototypical units against *all* features found in an image. Mutch et al. [35] employ matching of all 4000 higher-layer templates against features extracted in each pixel and scale of the resampled pyramid. This is also a drawback in layers of clustered histograms used in [1] and hierarchical classifiers in [27].

On the other hand, the success of hierarchical methods that do employ the principles of indexing and matching has been hindered by the use of hand-coded information. In [3], the authors use hand-crafted local edge features and only learn their global arrangements pertaining to specific object categories. The authors of [43] use predesigned filters and process the visual information in the feed-forward manner, while their recent version [46] exchanged the intermediate layer with random combinations of local edge arrangements rather than choosing the features in accordance with the natural statistics.

Approaches that do build the layers by learning and are able to make a sufficient number of them (by starting with simple features) mostly design the parts by histogramming the local neighborhoods of parts of the previous layers [1] or by learning the neural weights based on the responses on previous layers [27, 22]. Besides lacking the means of indexing, additional inherent limitation of such methods is the inefficiency in performing incremental learning; as the novel categories arrive, the whole hierarchy has to be re-adapted. Moreover, histograms do not enable robust top-down matching, while convolutional networks would have problems with the objects or features that are supersets/subsets of other features.

While the concepts of hierarchical representations, indexing and matching, statistical learning and incrementality have already been explored in the literature, to the best of our knowledge, they have not been part of a unifying framework. This chapter summarizes our recent, novel approach to building a hierarchical representation that aims to enable recognition and detection of a large number of object categories. Inspired by the principles of efficient indexing (*bottom-up*), robust matching (*top-down*), and ideas of compositionality, our approach *learns* a hierarchy of spatially flexible compositions, i.e., parts, in a completely unsupervised, statistics-driven manner. As the proposed architecture does not yet perform large-scale recognition, it makes important steps

towards scalable representations of visual categories.

The learning algorithm proposed in [21], which acquires a hierarchy of local edge arrangements by correlation, is in concept similar to our learning method. However, the approach demands registered training images, employs the use of a fixed grid, and is more concerned with the coarse-to-fine search of a particular category (i.e. faces) rather than finding features shared by many object classes.

4 Learning a Compositional Hierarchy of Parts

This Section summarizes our recently proposed framework that learns a hierarchical compositional representation of object structure from a set of natural images without supervision. The complete architecture addresses three major issues: the representation, learning the representation, and matching the representation against images.

The proposed representation takes the form of a compositional hierarchy with discrete nodes — compositions, also termed *parts*. Each part in the hierarchical vocabulary models loose spatial relations between its components, which at the lowest level correspond to simple contour fragments. The first layer of the hierarchy is fixed (but can in fact be a set of arbitrary filters) and it is also the only layer that operates directly on images. All the higher layers that make inference at subsequent stages are learned without supervision.

The approach is in essence composed of two recursively iterated steps:

- a layer-learning process that statistically extracts parts by sequentially increasing the number of subparts contained in local image neighborhoods, and
- a part matching step that finds the learned compositions in images with an efficient and robust *indexing and matching* scheme.

Layers are learned sequentially, layer after layer, optimally adjusting to the visual data. The advantage of the proposed learned representation lies in the capability to model exponential variability present in images, yet still retaining the computational efficiency by keeping the number of indexing links per each part approximately constant across layers.

The compositional representation and its envisioned properties are explained in Subsection 4.1. In Subsection 4.2 the hierarchical recognition process that matches the representation against images is discussed. We summarize the unsupervised learning procedure that extract a hierarchy of progressively more complex contour compositions in Subsection 4.3. Finally, Subsection 4.5 discusses a potential step towards categorical representations.

4.1 The compositional library of parts

We first present the properties of the envisioned hierarchical representation and the information that we would like to be coded within its discrete nodes — parts/compositions.

To abbreviate notation, let $\mathcal{L}_n$ denote the n-th Layer of the hierarchical library. We define the parts within the hierarchy recursively in the following way. Each part in $\mathcal{L}_n$ codes spatial relations between its constituent subparts

from a layer below. Formally, each composite part $\mathcal{P}_\ell^n$ in $\mathcal{L}_n$ is characterized by a *central subpart* and a list of remaining subparts with their positions relative to the center:

$$\mathcal{P}_\ell^n = \left(\mathcal{P}_{central}^{n-1}, \{\left(\mathcal{P}_j^{n-1}, \boldsymbol{\mu}_j, \Sigma_j\right)\}_j\right),$$

where $\boldsymbol{\mu}_j = (x_j, y_j)$ denotes the relative position of subpart $\mathcal{P}_j^{n-1}$, while Σ_j denotes the allowed variance of its position around (x_j, y_j). An example of a $\mathcal{L}_3$ composition is depicted in Figure 2.

The hierarchy starts with a fixed $\mathcal{L}_1$ composed of a set of arbitrary filters. Here we choose a set of Gabor filters that best respond to oriented edges.

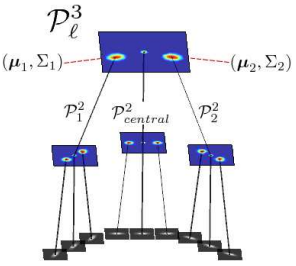

Figure 2: Example of a $\mathcal{L}_3$ composition.

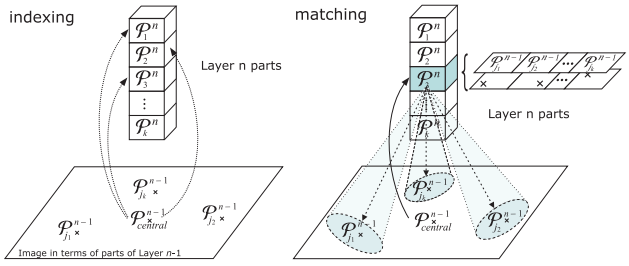

Figure 3: Left: Indexing – evoking higher level composite hypotheses. Right: Matching – verification of a composite library part.

4.2 Hierarchical recognition: The indexing and matching scheme

Let us suppose that the representation as described in the previous Subsection has already been acquired (how this is done will be explained in the following Subsection). This Subsection discusses how the incoming images are preprocessed and how the hierarchical representation is subsequently matched against the data. The complete procedure can be briefly summarized as follows:

- Each image is first processed with filters comprising $\mathcal{L}_1$ in order to get a (discrete) set of local contour fragments. Each contour segment, i.e. part, also codes its orientation and position in an image.

- Around each detected part, higher-order compositions are matched within the so called *indexing and matching scheme*. At each processed layer a discrete set of parts coding the types of the detected local structures and the corresponding locations is passed onto sequential matching stages. The hierarchical processing steps are all general in their traversal from one layer to the next and will thus be described in their general form. The procedure is illustrated in Figure 4.2.

- To attain robustness with respect to the scale of the objects (or their smaller substructures), the hierarchical recognition procedure is performed at several re-scaled versions of the image. This is schematically depicted in Figure 4.2.

Processing with Layer 1. For a given image, we first apply a set of $\mathcal{L}_1$ filters, here chosen to be the (odd and even) Gabor filters. Next, local maxima of the Gabor energy function [17] that are above a low threshold are found (these pertain to local oriented edges). The set of points (corresponding to the local maxima) together with their locations and labels (types) of filters that locally produced the maximal responses defines the list of $\mathcal{L}_1$ part detections, namely $\{\pi_k^1\}_k$. In general, π_k^n stands for a *realization* of the $\mathcal{L}_n$ part $\mathcal{P}_{\ell_k}^n$ with a corresponding location at which it was detected in an image; $\pi_k^n = \{\mathcal{P}_{\ell_k}^n, x_k, y_k\}$ (here k denotes the successive number of the found part).

This process is repeated at several image scales. However, for simplicity of notation we omit the delineation of part detections into separate scales. The obtained list of binary part detections, i.e. $\{\pi_k^1\}_k$, serves as input to subsequent hierarchical matching stages.

Hierarchical recognition. Let $\{\pi_k^{n-1}\}_k$ denote the list of the binary part detections from layer $\mathcal{L}_{n-1}$. In order to find a higher level image interpretation, the local neighborhoods around each detected π_k^{n-1} part are compared against the composite $\mathcal{L}_n$-parts stored in the hierarchical library. Each part realization $\pi_k^{n-1} = (\mathcal{P}_{\ell_k}^{n-1}, x_k, y_k)$ in the image under consideration is subjected to the *indexing and matching procedure* — efficient local search for higher level compositions.

The part $\mathcal{P}_{\ell_k}^{n-1}$ encoded in π_k^{n-1} plays the role of the central part in only a subset of all compositions at layer $\mathcal{L}_n$ of the library. This list is an internal part of the library and can be accessed in constant time during the online processing of images — the process referred to as *indexing*. The *matching* step demands comparing the local spatial neighborhood of π_k^{n-1} against the allowable (retrieved in the indexing step) prototypical compositions within the hierarchical library. Matching of one such composition, e.g. $\mathcal{P}_\ell^n = \big(\mathcal{P}_{central}^{n-1}, \{\big(\mathcal{P}_j^{n-1}, \boldsymbol{\mu}_j, \Sigma_j\big)\}_j\big)$ (where $\mathcal{P}_{central}^{n-1}$ corresponds to the part label $\mathcal{P}_{\ell_k}^{n-1}$), demands checking for the presence of all subparts $\{\big(\mathcal{P}_j^{n-1}, \boldsymbol{\mu}_j, \Sigma_j\big)\}_j$ pertaining to the composition $\mathcal{P}_\ell^n$ at their relative locations, $\boldsymbol{\mu}_j = (x_j, y_j)$, and positioned within the allowed variances, Σ_j, with respect to the position of the central part type $\mathcal{P}_{\ell_k}^{n-1}$ coded in π_k^{n-1}. The indexing and matching procedure is schematically depicted in Figure 3.

4.3 Unsupervised learning of part compositions

The basic idea behind the learning procedure is to extract statistically salient compositions that encode spatial relations between the constituent parts from the layer below. Each modeled relation between components allows also for some displacement (variance) in spatial position.

The learning algorithm is in principle general – proceeding in the same manner when building each additional layer. It will thus be described in its general form.

The learning process consists of three stages, namely, (1) the local inhibition performed around each image feature (part), followed by (2) the statistical updating of the so-called *spatial maps* that capture pairwise geometric relations between parts, and finally, (3) learning the higher order compositions by tracking co-occurrence of spatial pairs. We must emphasize that each final composition can have a varying number of subcomponents (the number can be anything from

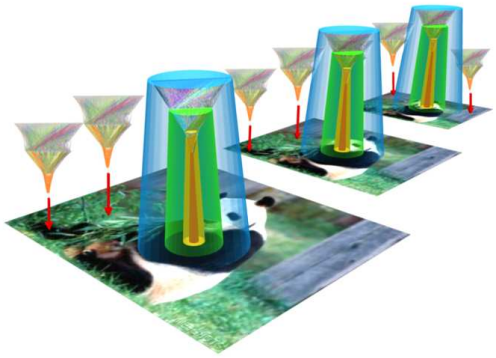

Figure 4: The learned hierarchical library of parts is applied in each image point (robustness to position of objects) and several image scales (robustness to objects' size)

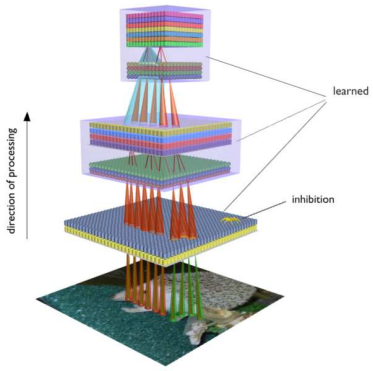

Figure 5: The hierarchical recognition architecture.

2 and larger).

Learning is performed by gathering statistics over a large body of natural images processed up to the last (learned) layer in the hierarchical library, e.g. $\mathcal{L}_{n-1}$. Each image is thus represented by a list of parts with corresponding locations, $\{\pi_k^{n-1}\}_k$. A small local neighborhood around each π_k^{n-1} will be inspected in a two-stage process. The first, most crucial step aims to reduce the unnecessary redundancy coded in neighboring parts, referred to as *local inhibition*. Since each π_k^{n-1} is an $(n-1)$-th order composition, it is in fact a set union of a few detected $\mathcal{L}_1$ parts. Within the inhibition step we remove all neighboring parts around π_k^{n-1} that have a large set intersection with respect to the $\mathcal{L}_1$ image parts. This step removes all features that code a large portion of edge structure already coded by π_k^{n-1}. In the next step, learning is performed by tracking frequent co-occurrences of part types and their relative locations.

The learning process commends by forming a set of all allowable pairs of part identities. The list is accompanied by a set of empty matrices, where the dimensions correspond to the spatial extent of the local neighborhoods. The prepared set thus contains information of type: $\mathcal{C}_{k,j}^n := (\mathcal{P}_k^{n-1}, \mathcal{P}_j^{n-1}, \mathcal{V}_{k,j})$, where $\mathcal{V}_{k,j}$ represents a local spatial voting space for the corresponding combination of pairs of part types $\mathcal{P}_k^{n-1}$ and $\mathcal{P}_j^{n-1}$.

Structure of small neighborhoods in terms of part locations is inspected around each part, π_k^{n-1}. The philosophy of local receptive field processing is the following: the location of each part $\pi_j^{n-1} = (\mathcal{P}_j^{n-1}, x_j, y_j)$ within the neighborhood of and relative to π_k^{n-1} will update the voting space $\mathcal{V}_{k,j}$ in $\mathcal{C}_{k,j}$ accordingly:

$$x := cx + x_j - x_k, \;\; y := cy + y_j - y_k$$
$$\mathcal{V}_{k,j}(x,y) = \mathcal{V}_{k,j}(x,y) + 1,$$

where (cx, cy) denotes the center of the spatial map $\mathcal{V}_{k,j}$.

After all images are processed, we detect voting peaks in the learned spatial maps $\mathcal{V}_{k,j}$, and for each peak, a spatial surrounding area is formed – modeled by a Gaussian distribution, $(\boldsymbol{\mu}_j, \Sigma_j)$.

In the final step, the local image neighborhoods are checked once again by projecting the learned spatial pairs and repeating the learning process by increasing the number of subparts modeled in a composition (the reader is referred to [19] for details) or by tracking the most frequent co-occurrences of the projected spatial pairs.

The final selection of composite parts follows the indexibility constraint, i.e., each part of the lower, $(n-1)$-th Layer, must not index into too many higher layer compositions. Thus the compositions acquired in the learning procedure are sorted according to their decreasing probabilities and only a number of statistically most salient compositions consequently define the next layer. We set the upper bound to the order of $10-20$ times the number of parts in the previous, $(n-1)$-th Layer, meaning that on average each part in $\mathcal{L}_{n-1}$ indexes into 10 to 20 composite parts in $\mathcal{L}_n$. The thresholds used are chosen to comply with the available computational resources and affect only the number of finally selected parts and therefore the efficiency of the representation.

4.4 Grouping of part labels by similarity and co-occurrence

The problem with the learned compositions is the fact that they are realized as discrete labels (part types) without a proper geometrical parametrization that would enable a comparison between them. Consequently, two visually similar curvatures are likely to be encoded in two different hierarchical compositions. We deal with this issue in two ways. One approach is grouping by co-occurrence [19] where parts that frequently co-occur in close spatial proximity of one another are assigned the same label (part type).

However, two visual shapes that are only similar to a certain extent are likely to have a small, random co-occurrence. It is thus crucial to also have the means of comparing two different parts in a geometrical sense. Since each part is formed as a recursive spatially loose composition, a comparison can be performed in a similar manner. We consider two parts to be perceptually similar if both have a similar spatial configuration of subparts. For details of such recursive comparison of compositions we refer the reader to our original work [18].

4.5 Category-specific higher layers

Learning the lower-layer sharable parts in a category-independent way can only get so far — the overall statistical significance drops, while the number of parts reaches its critical value for learning. Thus, learning of higher layers proceeds only on a subset of parts — the ones that are the most repeatable for a specific category. Specifically, the learning of higher layers is performed in images of individual categories, whereby the final categorical layer then combines the most repeatable parts through the object center to form the representation of a category.

5 Experimental results

We applied our method to a collection of 1500 natural images containing a number of diverse categories (cars, faces, mugs, dogs, etc.). A few examples

of the images used for learning are presented in Figure 6. The complete learning process took approximately 5 hours on one core of an Intel Core-2 CPU 2.4 Ghz computer. The learning procedure produced a compositional hierarchy consisting of 160 parts on Layer 2 and 553 parts on Layer 3 (a few examples from both layers are depicted in Figure 7). The learned features include corners, end-stopped lines, various curvatures, T- and L-junctions, etc. It must be noted that a much smaller set of images is in fact needed to result in virtually the same hierarchy. Experiments have shown that learning on approximately 50 images produces almost exactly the same $\mathcal{L}_2$ vocabulary as the larger set of 1500 images, while approximately 200 images are required to learn the third layer of the hierarchy. We have also experimented with using different filters at $\mathcal{L}_1$. The learned vocabulary for the Gabor filters that also take into account the polarity of edges is presented in Figure 8. We must emphasize, however, that both figures 7 and 8 only show the contours that the parts produce maximal responses to. Each learned part in the vocabulary is in fact a composition of the form shown in Figure 2.

To put the proposed hierarchical framework in relation to other hierarchical approaches as well as other categorization methods, which focus primarily on shape information, the approach was tested on the Caltech 101 database [15]. The Caltech 101 dataset contains images of 101 different object categories with the additional background category. The number of images varies from 31 to 800 per category, with the average image size of roughly 300×300 pixels. Each image was processed on 3 different scales, spaced apart by $\sqrt{2}$. The average processing times per image per layer (including all three scales) obtained in a C++ implementation are the following: 1.6 seconds for $\mathcal{L}_1$, 0.54 seconds for $\mathcal{L}_2$ and 0.66 seconds for $\mathcal{L}_3$. The features were combined with a linear SVM for multiclass classification. For both, 15 and 30 images for training we obtained 60.5% and 66.5% classification accuracy, respectively, which is the best result reported by a hierarchical approach so-far. While we do not believe that SVM classification is the proper form of categorization, the experiments were performed to demonstrate the utility of the learned features with respect to the features used in the current state-of-the-art approaches applied in similar classification settings.

We have also attempted learning higher — categorical layers, using images of specific categories for training. The learning of categorical layers, namely $\mathcal{L}_4$, was run only on images containing faces (6 out of 20 images used for training are shown in the top row of Figure 9), cars (6 out of 20 training images are depicted in the middle row of Figure 9), and mugs (all training images are presented in the bottom row of Figure 9). The obtained parts were then learned relative to centers of faces, cars and mugs, respectively, to produce $\mathcal{L}_5$ - *category layer*. Figure 10 shows the learned layers, while Figure 11 depicts the learned hierarchical vocabulary for faces with compositional links shown (second image in the top row). It must be noted that the first three layers in the hierarchy are general — the same for faces as for cars and mugs, while only layers 4 and 5 are not sharable among the three categories. The recognition and the subsequent segmentation (tracing the recognition nodes down the the image) of parts through the hierarchy on example images is presented in Figure 11. In Figure 12 several examples of mug detections are presented, showing the approach is capable of recognizing various class members, hand-drawn objects as well as peculiar mug-like compositions (a drawn handle plus a basket, a drawn

glass plus some handle-like object added accordingly).

Figure 6: Examples of natural images used to learn the category-independent layers.

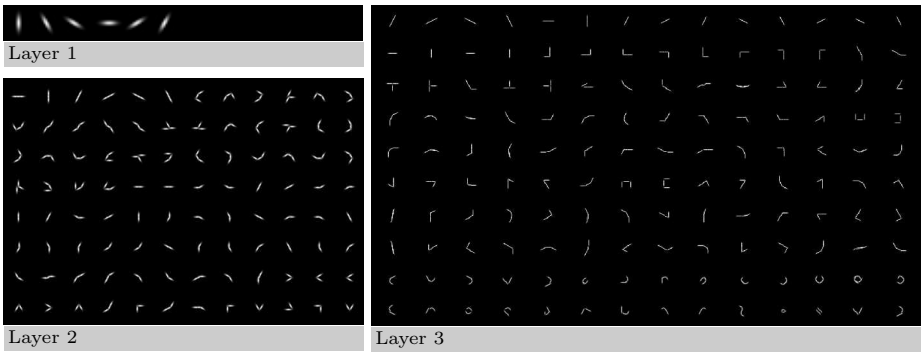

Figure 7: $\mathcal{L}_1$ (fixed - oriented edges), and learned $\mathcal{L}_2$ and $\mathcal{L}_3$ parts (only a subset is shown) used in the Caltech 101 experiments.

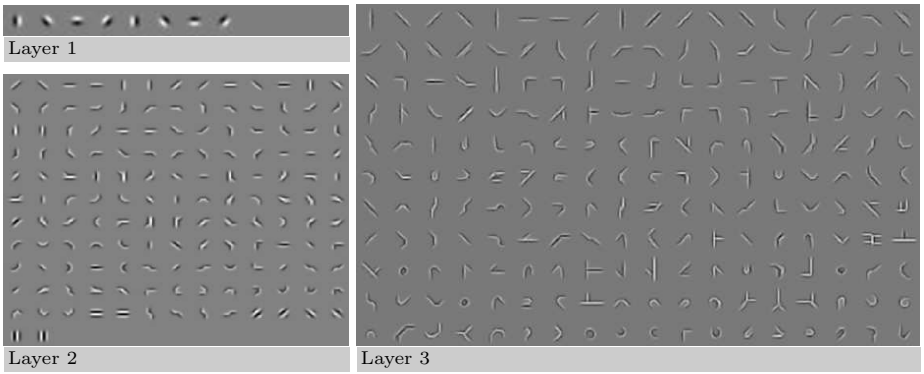

Figure 8: $\mathcal{L}_1$ (fixed - polarity filters), and learned $\mathcal{L}_2$ and $\mathcal{L}_3$ parts (only a subset is shown).

Figure 9: Examples of images used to learn the categorical layers.

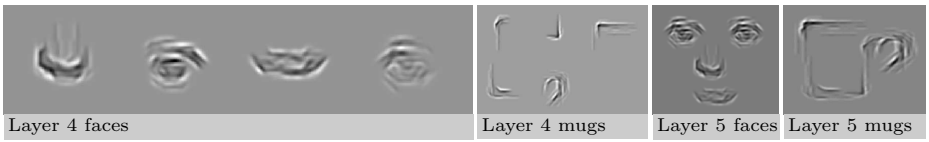

Figure 10: Learned categorical layers $\mathcal{L}_4$ and $\mathcal{L}_5$ for faces and mugs.

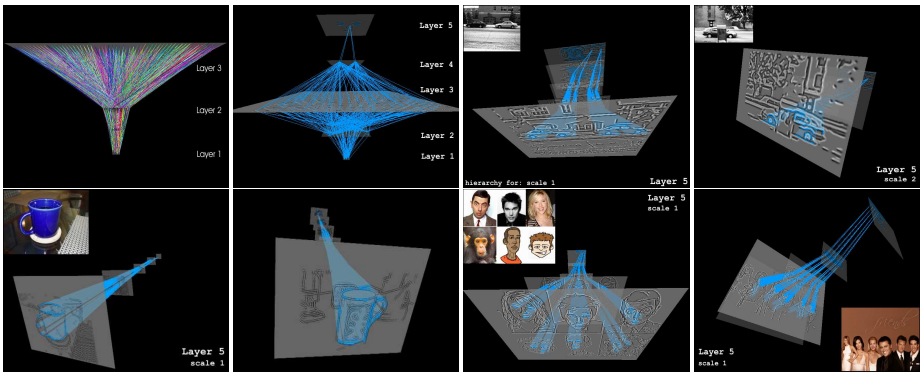

Figure 11: Top left two images: learned 3−layer hierarchy for the Caltech experiment; learned hierarchy for faces with compositional links shown. Examples of detections of categories cars, mugs, and faces, where the first three layers in the library are common to all three categories.

6 Summary and Discussion

This chapter summarized our recent approach to building a representation of object structure. The method learns a hierarchy of flexible compositions in an unsupervised manner in lower, category-independent layers, while requiring minimal supervision to learn higher, categorical layers.

Furthermore, the design of parts is incremental, where new categories can be continuously added to the hierarchy. Since the hierarchy is built as an efficient

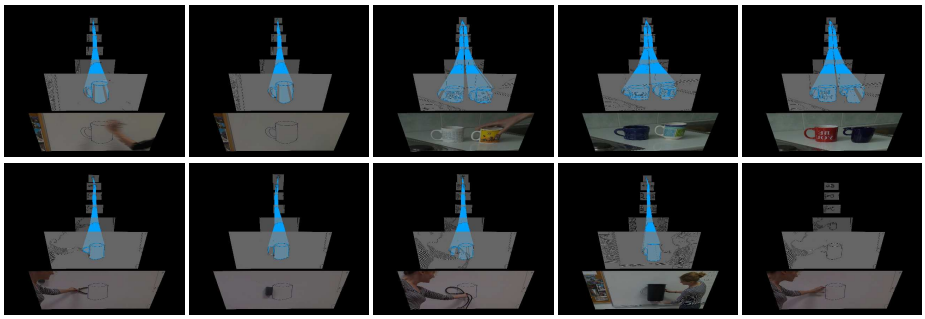

Figure 12: Detections of mugs.

indexing machine, the system can computationally handle an exponentially increasing number of parts with each additional layer. The results show that only a small number of higher layer parts are needed to represent individual categories, thus the proposed scheme would potentially allow for an efficient representation of a large number of visual categories.

Our future work includes improvements over the current creation of the categorical layers, adding different modalities such as color, texture and motion, and improving inference by using iterated loops between the bottom-up and top-down information flow.

Acknowledgemnts

This research has been supported in part by the following funds: Research program Computer Vision P2-0214 (Slovenian Ministry of Higher Education, Science and Technology), EU FP6-004250-IP project CoSy, EU FP6-511051 project MOBVIS, and EU FP7-215843 project POETICON.

References

[1] A. Agarwal and B. Triggs. Hyperfeatures - multilevel local coding for visual recognition. In *ECCV (1)*, pages 30–43, 2006. 2, 7

[2] Y. Amit. *2d Object Detection and Recognition: Models, Algorithms and Networks*. MIT Press, Cambridge, 2002. 1, 2, 7

[3] Y. Amit and D. Geman. A computational model for visual selection. *Neural Comp.*, 11(7):1691–1715, 1999. 1, 2, 3, 7

[4] A. Anzai, X. Peng, and D. C. Van Essen. Neurons in monkey visual areaa v2 encode combinations of orientations. *Nature Neuroscience*, 10(10):1313–1321, 2007. 2, 5

[5] H. B. Barlow. Cerebral cortex as a model builder. pages 37–46, 1985. 6

[6] E. Bienenstock. Composition. In A. Aertsen and V. Braitenberg, editors, *Brain Theory - Biological Basis and Computational Theory of Vision*, pages 269–300. Elsevier, 1996. 3

[7] E. Bienenstock and S. Geman. Compositionality in neural systems. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 223–226. MIT Press, 1995. 2, 3, 5

[8] A. Califano and R. Mohan. Multidimensional indexing for recognizing visual shapes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(4):373–392, 1994. 1, 2, 3, 7

[9] C. E. Connor, S. L. Brincat, and A. Pasupathy. Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology*, 17(2):140–147, 2007. 1, 2, 5

[10] S. Dickinson. The evolution of object categorization and the challenge of image abstraction. In S. Dickinson, A. Leonardis, B. Schiele, and M. J. Tarr, editors, *Object Categorization: Computer and Human Vision Perspectives*. Springer-Verlag, 2008. 1

[11] S. Dickinson, A. Pentland, and A. Rosenfeld. 3-D shape recovery using distributed aspect matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(2):174–198, 1992. 1, 2

[12] S. Edelman and N. Intrator. Unsupervised statistical learning in vision: computational principles, biological evidence. In *extended abstract of invited talk at the ECCV-2004 Workshop on Statistical Learning in Computer Vision*, 2004. 6

[13] G. J. Ettinger. Hierarchical object recognition using libraries of parameterized model sub-parts. Technical report, MIT, 1987. 1, 2, 3

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 1

[15] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *IEEE CVPR'04, Workshop on Generative-Model Based Vision*, 2004. 1, 13

[16] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(1):36–51, 2008. 3

[17] S. Fidler, G. Berginc, and A. Leonardis. Hierarchical statistical learning of generic parts of object structure. In *CVPR*, pages 182–189, 2006. 1, 6, 10

[18] S. Fidler, M. Boben, and A. Leonardis. Similarity-based cross-layered hierarchical representation for object categorization. In *CVPR'08*, 2008. 12

[19] S. Fidler and A. Leonardis. Towards scalable representations of visual categories: Learning a hierarchy of parts. In *CVPR'07*, 2007. 1, 3, 6, 12

[20] J. Fiser and R. N. Aslin. Statistical learning of new visual feature combinations by infants. *Proc Natl Acad Sci U S A*, 99(24):15822–15826, 2002. 6

[21] F. Fleuret and D. Geman. Coarse-to-fine face detection. *IJCV*, 41(1/2):85–107, 2001. 2, 8

[22] K. Fukushima, S. Miyake, and T. Ito. Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Systems, Man and Cybernetics*, 13(3):826–834, 1983. 2, 7

[23] S. Geman, D. Potter, and Z. Chi. Composition systems. *Quarterly of Applied Mathematics*, 60(4):707–736, 2002. 1, 2, 3

[24] W. E. L. Grimson and T. Lozano-Perez. Localizing overlapping parts by searching the interpretation tree. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):469–482, 1987. 1

[25] J. Hawkins and S. Blakeslee. *On Intelligence*. Times Books, 2004. 2, 6

[26] G. E. Hinton. Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11(10):428–434, 2007. 1

[27] F.-J. Huang and Y. LeCun. Large-scale learning with svm and convolutional nets for generic object categorization. In *CVPR'06*, pages 284–291, 2006. 5, 7

[28] J. Hup, A. James, B. Payne, S. Lomber, P. Girard, and J. Bullier. Cortical feedback improves discrimination between figure and background by v1, v2 and v3 neurons. *Nature*, 394:784–787, 1998. 4

[29] M. Ito and H. Komatsu. Representation of angles embedded within contour stimuli in area v2 of macaque monkeys. *The Journal of Neuroscience*, 24(13):3313–3324, 2004. 5

[30] S. Krempp, D. Geman, and Y. Amit. Sequential learning of reusable parts for object detection. Technical report, 2002. 2

[31] T. S. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis*, 20(7):1434–1448, 2003. 4

[32] G. Medioni, T. Fan, and R. Nevatia. Recognizing 3-d objects using surface descriptions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2(11):1140–1157, 1989. 1

[33] B. W. Mel and J. Fiser. Minimizing binding errors using learned conjunctive features. *Neural Computation*, 12(4):731–762, 2000. 1

[34] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *CVPR'06*, pages 26–36, 2006. 2, 7

[35] J. Mutch and D. G. Lowe. Multiclass object recognition with sparse, localized features. In *CVPR06*, pages 11–18, 2006. 7

[36] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE CVPR*, volume 2, pages 2161–2168, 2006. 2

[37] B. Ommer and J. M. Buhmann. Learning compositional categorization models. In *ECCV'06*, pages 316–329, 2006. 3

[38] B. Ommer and J. M. Buhmann. Learning the compositional nature of visual objects. In *CVPR'07*, 2007. 2

[39] H. Op de Beeck, J. Haushofer, and N. Kanwisher. Interpreting fmri data: maps, modules, and dimensions. *Nature Reviews Neuroscience*, 9:123–135, 2008. 2

[40] A. Pasupathy and C. Connor. Responses to contour features in macaque area v4. *Journal of Neurophysiology*, 82(5):2490–2502, 1999. 5

[41] N. Pinto, D. Cox, and J. DiCarlo. Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4(1):151–156, 2008. 1

[42] M. A. Ranzato, F.-J. Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR'07*, 2007. 2

[43] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999. 5, 7

[44] E. T. Rolls and G. Deco. *Computational Neuroscience of Vision.* Oxford Univ. Press, 2002. 1, 6

[45] F. Scalzo and J. H. Piater. Statistical learning of visual feature hierarchies. In *Workshop on Learning, CVPR*, 2005. 2

[46] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Object recognition with cortex-like mechanisms. *PAMI*, 29(3):411–426, 2007. 2, 7

[47] A. Shokoufandeh, L. Bretzner, D. Macrini, M. Demirci, C. Jonsson, and S. Dickinson. The representation and matching of categorical shape. *Computer Vision and Image Understanding*, Vol. 103:139–154, 2006. 1, 2, 3

[48] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV'05*, pages 1331–1338, 2005. 2, 7

[49] S. J. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996. 4

[50] S. Todorovic and N. Ahuja. Unsupervised category modeling, recognition, and segmentation in images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2007. 2, 3

[51] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. 2

[52] J. Tsotsos. What roles can attention play in recognition? In *7th International Conference on Development and Learning*, 2008. 4

[53] J. K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13(3):423–469, 1990. 1

[54] K. Tsunoda, Y. Yamane, M. Nishizaki, and M. Tanifuji. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, (4):832–838, 2001. 2, 5

[55] S. Ullman. Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn. Sci.*, 11:58–64, 2007. 2, 4

[56] S. Ullman and B. Epshtein. *Visual Classification by a Hierarchy of Extended Features.* Towards Category-Level Object Recognition. Springer-Verlag, 2006. 2, 7

[57] R. VanRullen. The power of the feed-forward sweep. *Advances in Cognitive Psychology*, 3(1–2):167–176, 2007. 4

[58] L. Zhu and A. Yuille. A hierarchical compositional system for rapid object detection. In *NIPS'05*, 2005. 3

[59] S. Zhu and D. Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2(4):259–362, 2006. 3