Int J Comput Vis (2011) 93: 117–140
DOI 10.1007/s11263-010-0391-1

Context, Computation, and Optimal ROC Performance in Hierarchical Models

Lo-Bin Chang · Ya Jin · Wei Zhang · Eran Borenstein · Stuart Geman

Received: 11 June 2010 / Accepted: 21 September 2010 / Published online: 9 October 2010

Abstract It is widely recognized that human vision relies on contextual information, typically arising from each of many levels of analysis. Local gradient information, otherwise ambiguous, is seen as part of a smooth contour or sharp angle in the context of an object's boundary or corner. A stroke or degraded letter, unreadable by itself, contributes to the perception of a familiar word in the context of the surrounding strokes and letters. The iconic Dalmatian dog stays invisible until a multitude of clues about body parts and posture, and figure and ground, are coherently integrated. Context is always based on knowledge about the composition of parts that make up a whole, as in the arrangement of strokes that make up a letter, the arrangement of body parts that make up an animal, or the poses and postures of individuals that make up a mob. From this point of view, the hierarchy of contextual information available to an observer derives from the compositional nature of the world being observed. We will formulate this combinatorial viewpoint in terms of probability distributions and examine the computational implications. Whereas optimal recognition performance in this formulation is NP-complete, we will give mathematical and experimental evidence that a properly orchestrated computational algorithm can achieve nearly optimal recognition within a feasible number of operations. We will interpret the notions of bottom-up and top-down processing as steps in the staging of one such orchestration.

Keywords Context · Hierarchy · Parts · Composition · ROC performance · Bottom-up processing · Top-down processing · Coarse-to-fine search

1 Introduction

A frame from a 1920's shot of the expressionless face of the Russian actor Ivan Mozzhukhin is shown, repeatedly, on the right hand side of Fig. 1. The shot was captured by the director Lev Kuleshov as part of an experiment context and a study of its role in the cinematic experience. In three separate clips, Kuleshov juxtaposes the shot with a dead child lying in an open coffin, a seductive actress, or a bowl of soup. Asked to interpret Mozzhukhin's expression, audiences reported sadness, lust, or hunger depending on whether the expression followed the images of the dead child, the seductive actress, or the bowl of soup. Many praised the actor's skill. The idea that the movie-going experience is based on composition as much as content became the basis for the so-called montage school of Russian cinema and it remains an essential tool of modern filmmaking.

The effects of context on human perception have been well studied for hundreds of years, and are well illustrated with many familiar illusions involving size and boundary perception, grouping, and shading. But most contextual ef-

Partially supported by the Office of Naval Research under N000140610749, and the National Science Foundation under ITR-0427223 and DMS-1007593.

L.-B. Chang · Y. Jin · W. Zhang · E. Borenstein · S. Geman (✉)
Division of Applied Mathematics, Brown University, Providence, RI, USA
e-mail: stuart_geman@brown.edu

L.-B. Chang
e-mail: lo-bin_chang@brown.edu

Y. Jin
e-mail: jin.ya76@gmail.com

W. Zhang
e-mail: weizhang.brown@gmail.com

E. Borenstein
e-mail: eran.borenstein@gmail.com

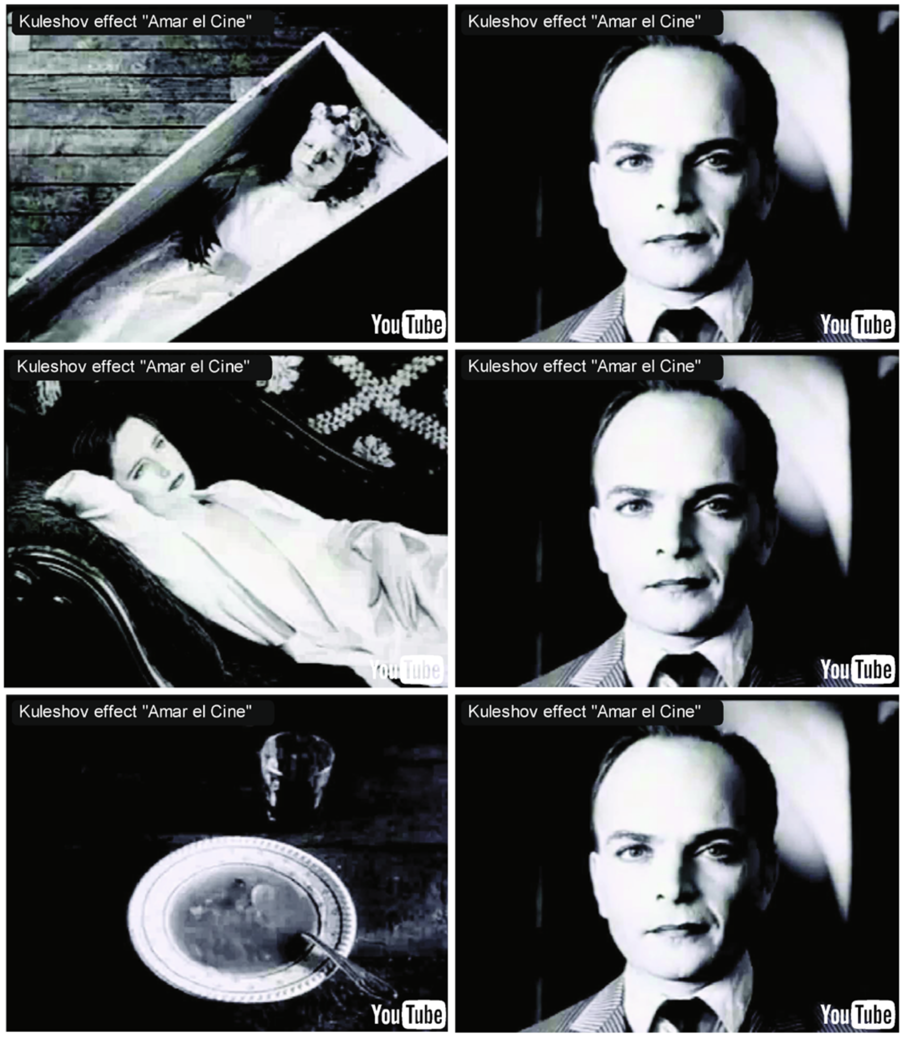

Kuleshov effect "Amar el Cine"
Kuleshov effect "Amar el Cine"
Kuleshov effect "Amar el Cine"
Kuleshov effect "Amar el Cine"
Kuleshov effect "Amar el Cine"
Kuleshov effect "Amar el Cine"
YouTube

Fig. 1 The Kuleshov Effect. *Top Row.* Frames from a sequence with a dead child followed by a shot of Ivan Mozzhukhin's face. *Middle Row.* Frames from a sequence with an actress in a seductive pose, followed by the same shot of Mozzhukhin's face. *Bottom Row.* Frames from a sequence with a bowl of soup, followed again by the same shot of Mozzhukhin's face. Audiences viewing the clips ascribe different emotions to the same expression, sadness, lust, or hunger, depending on the context

fects are not illusory. Sighted people are all experts at vision, which makes it difficult, if not impossible, to appreciate the multiple levels of context that critically influence virtually every visual perception. On the other hand, engineers trying to build artificial vision systems invariably discover the ambiguities in local raw pixel data. It is often impossible to decipher cursive, even one's own cursive, without a measure of context, which might come from any one of many levels, including topic, sentence, or just neighboring words or letters. The same effect is striking when applied to auditory signals, where, for example, words spliced from continuous speech are often unintelligible.

Many cognitive scientist would argue that the layers of context that influence the perception of a part or object (e.g. a phoneme or a word) are a manifestation of the compositional nature of mental representation (e.g. Fodor and Pylyshyn 1988). The vision scientist might be tempted to turn this around and say that these representations are themselves manifestations of the compositional nature of the visual or auditory world (cf. Warren 2010), but either way, or both, the evidence is that biological-level performance in perceptual tasks relies on knowledge about the relational groupings of parts into wholes, simultaneously at multiple levels of a hierarchy. This combinatorial, or compositional, viewpoint is a common starting point for discriminative or generative models of vision (Sudderth et al. 2005; Ommer and Buhmann 2007; Epshtein and Ullman 2005; Bengio and LeCun 2007; Amit and Trouvé 2007; Serre et al. 2007; Ahuja and Todorovic 2008), often within grammar or grammar-like organizations (Burl and Perona 1998; Zhu and Mumford 2006; Chen et al. 2007; Fidler and Leonardis 2007; Zhang 2009; Chang 2010; Felzenszwalb and McAllester 2010).

The idea in generative models is to use probability distributions to capture likely and unlikely arrangements, starting from arrangements of local features (e.g. local edges or

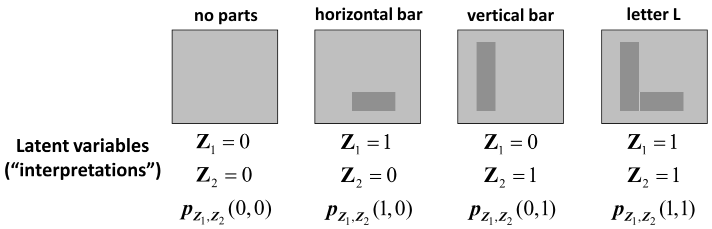

no parts
horizontal bar
vertical bar
letter L
Latent variables
("interpretations")
$Z_1 = 0$ $Z_2 = 0$ $p_{Z_1,Z_2}(0,0)$
$Z_1 = 1$ $Z_2 = 0$ $p_{Z_1,Z_2}(1,0)$
$Z_1 = 0$ $Z_2 = 1$ $p_{Z_1,Z_2}(0,1)$
$Z_1 = 1$ $Z_2 = 1$ $p_{Z_1,Z_2}(1,1)$

Fig. 2 **Minimal Compositional World.** There are only two parts, horizontal bar and vertical bar, and one object, the letter L. The presence of both parts is always interpreted as the letter L. In the experiments, $P_{Z_1,Z_2}(0,0) = 0.6$, $P_{Z_1,Z_2}(1,0) = P_{Z_1,Z_2}(0,1) = 0.1$, and $P_{Z_1,Z_2}(1,1) = 0.2$

texture elements), and in principle continuing recursively to high-level expert knowledge (e.g. a curator's knowledge about antique furniture, a grandmaster's knowledge about strengths and weaknesses in an arrangement of chess pieces). We will adopt the generative compositional viewpoint here, and use it to examine the practical problems of clutter, false alarms, and computational burden in artificial vision systems.

Clutter and the Limits of Artificial Vision Systems It is one thing to build a classification device that performs on images with single objects placed in simple backgrounds and quite another to find and classify these same objects in unconstrained scenes. Everyone who builds vision systems knows this. Real background has structure, and too often the structure masquerades as bits and pieces of the objects of interest. Run a correlator for an eye, with say a 10 × 10 patch, on backgrounds in an ensemble of images with bricks and trees and cars (e.g. mid-day Manhattan street scenes as captured by Google's Street View) and you'll probably get many good matches per frame, if "good match" is defined to be at least as good as 5% of the matches to real eyes in the same scenes. This kind of thing is to be expected, if you buy the compositional point of view. In particular, the parts of an object of interest, such as a face, are reusable and can be found among the pieces making up many other structures. It is not that there are actual eyes in and among the bricks, bark, and leaves, but that poorly-resolved oval shapes, with darker centers and lighter surrounds, are not uncommon and certainly not unique to faces. Indeed, if it were otherwise, then excellent performance on face detection tasks could be achieved by looking for nothing but eyes. But state-of-the-art face-detection algorithms, still not as good as human observers, require more than just eyes. Google Street View, in order to achieve a high certainty of detecting and obscuring real faces, blurs many false-detections on car wheels, trees, or just about anyplace that includes structured or textured background. When operating at the same detection level, humans get almost no false positives.

In general, artificial vision systems operating at the high-detection end of the ROC curve suffer many more false detections in unconstrained scenes than do human observers. If we think of a "part" as being *defined by* its local appearance, rather than its participation in any particular object, then we can think of these false detections as typically arising from an unlucky arrangement of parts of the objects of interest. A human interprets these same arrangements for what they are: parts of other objects, or objects in their own right. One could reasonably argue, then, that solving one vision problem, say the detection of a single object, requires solving many vision problems, at least the detection of any other object that shares aspects of its appearance, i.e. shares parts, with the object of interest. How much knowledge is needed to achieve biological-level performance on a single vision task? Is it necessary to know about all objects to accurately detect a single object? Is vision "AI-complete"?

We will argue in the opposite direction. We will give evidence that, to the extent the world is compositional, a vision system can achieve nearly optimal performance on a particular vision task, involving a single selected object or a particular library of objects, by modeling only the object or objects of interest. The idea is that most false detections occur at background locations that share bits and pieces of the objects of interest, suggesting that the objects themselves, viewed as compositional, define adequate background models through their own subparts and arrangements of subparts; in a compositional world, objects define their own background models (Jin and Geman 2006).

Matching Templates Versus Matching Parts We often think of cascades and other coarse-to-fine strategies as computational imperatives. Even if we had a full-blown model for the appearance of an object, it would not be feasible to search for it at every pose (already six degrees of freedom for a rigid object). Except in very special circumstances, practical vision systems have to use some form of coarse-to-fine search. This usually involves a very simple first pass that highlights candidate poses, followed by a sequence of more refined and constrained searches in the neighborhoods of the candidate poses. Computation might be organized as a tree, for example to search simultaneously for multiple objects or to postpone decisions about pose or identity by exploring multiple branches, or as a cascade, which might be suitable for single objects and limited pose variation. The computational advantages are well documented, both from a practical and a theoretical standpoint (Fleuret and Geman 2001; Viola and Jones 2001; Moreels and Perona 2008; Blanchard and Geman 2005; Amit and Trouvé 2010).

But computation might not be the whole story. There might be other reasons for preferring a divide-and-conquer strategy. Consider an imaginary object $\mathcal{O}$ that can appear at only one pose, and an imaginary situation in which we have a fully specified render model for the distribution on images given that $\mathcal{O}$ is present. How would we test for $\mathcal{O}$? How do we compare the hypothesis "$\mathcal{O}$ is present" to the alternative "$\mathcal{O}$ is absent"? It is not enough to have an appearance model for $\mathcal{O}$; we also need an appearance model for scenes without $\mathcal{O}$. The trouble is that "$\mathcal{O}$ absent" is an unimaginably large mixture. What is more, as we have already observed, this mixture will typically include components that represent objects with similarities to $\mathcal{O}$, portions of which might be essentially indistinguishable from portions of $\mathcal{O}$.

An expedient approach would be to adopt a simple "background model," meaning some kind of manageable alternative distribution, such as iid pixel intensities or more generally a random field that might capture local correlations. To the extent that the background model is accurate, the likelihood ratio, the ratio of the probability of the observed image given that $\mathcal{O}$ is present to the probability under the background model, is an optimal statistic for this two-class problem (i.e. thresholding on the ratio will minimize false alarms at any given detection rate). Another approach, also expedient, would be to test for the presence of parts of $\mathcal{O}$. If all of the parts are found, then declare that $\mathcal{O}$ is present. The same simple background model could be used, locally, to test for the individual parts.

Both approaches have advantages. The first, which is essentially a template match, is relatively robust to a noisy presentation of the object. The parts may be difficult to confirm, individually, but the collective evidence could be strong. The second, although vulnerable to a poorly rendered part, has an easier time distinguishing false alarms when the actual scene contains parts and objects that resemble pieces of $\mathcal{O}$, but not $\mathcal{O}$ itself. Our purpose is to argue, through mathematical and empirical evidence, that the second approach, parts-based testing, is superior, especially when operating at a high-detection threshold. In fact, it might not be far from optimal. We will propose a particular version of parts-based testing, similar to a particular scheduling of belief propagation, that is suitable for compositional models and is recursive for hierarchical models. We will interpret the computational steps in terms of context as well as the notions of "bottom-up" and "top-down" processing.

Parts-based testing can be viewed as an example of coarse-to-fine search. The presence of a part does not confirm the presence of an object, but it does narrow the hypothesis space. And if the part is essential, then the search can be abandoned early when the part is missing. We will discuss connections to sequential testing and potential computational efficiencies, as well as the challenges that go with the typical case in which multiple parts can have multiple poses, defining an exponential number of instantiations of an object.

We begin, in Sect. 2, with a simple thought experiment, not unlike the discussion here of the fictional object $\mathcal{O}$. We will formulate the detection problem in such a way that we can compare three approaches, the optimal approach (based on the Neyman-Pearson Lemma), the template approach, and the parts-based approach. The comparison will be mathematical, via comparisons of the area under the ROC curve for each of the three alternatives, and via experiments with simulated and real data, chosen to be simple enough that good approximations to each of the three approaches can be computed. Section 3 is devoted to a discussion of generalizations of the results in Sect. 2, some of which are proven and some of which are speculative. In Sect. 4 we focus on hierarchical generative models and propose a recursive parts-based approach to orchestrating computations. Throughout, our examples are chosen, if not actually rigged, to make possible the comparison to optimal detection. We can only speculate that these comparisons will remain valid when operating on more complex compositional models. Section 5 contains a summary and some concluding remarks.

2 A Simple World of Parts and Objects

We start with a minimal world of parts and objects, depicted in Fig. 2. There are two parts, vertical and horizontal bars, and one object, the letter L. The model is generative and includes latent variables, one for each part, that define an "interpretation," and a conditional rendering distribution for the image given an interpretation. The latent variables, denoted Z_1 and Z_2 for the horizontal and vertical bars, respectively, are each binary ($Z_1, Z_2 \in \{0, 1\}$), representing the absence (0) or presence (1) of a part. The joint probability on parts is $P(Z_1 = z_1, Z_2 = z_2)$, $z_1, z_2 \in \{0, 1\}$. Referring to Fig. 3, the set of all pixels is denoted S and the subsets of pixels affected by the presence or absence of the parts are denoted S^1 and S^2, for the horizontal and vertical bars respectively. We will refer to S^1 and S^2 as the "supports" of their respective parts. The intensity of pixel $s \in S$ is treated as a random variable and is denoted X_s. Generically, given any set of pixels $A \subseteq S$, we use lexicographic (raster) ordering to define a vector of intensities X_A from the set $\{X_s : s \in A\}$.

The generative model generates an image (x_S) by first generating an interpretation (z_1, z_2) according to the joint

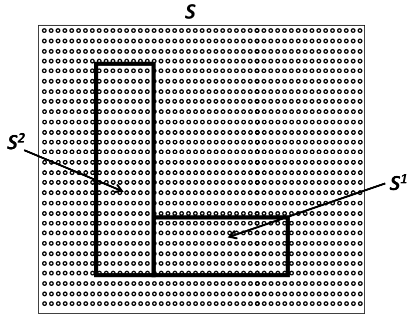


Fig. 3 **Pixel Lattice.** S is the set of pixels. $S^1 \subseteq S$ (the "support of part 1") and $S^2 \subseteq S$ (the "support of part 2") are the subsets of pixels at which horizontal and vertical bars appear, respectively

distribution on Z_1 and Z_2; then assigning intensities iid in S^1 and, independently, iid in S^2 according to $N(z_1, 1)$ and $N(z_2, 1)$, respectively; and finally, independently of everything else, assigning intensities iid in $S \setminus (S^1 \cup S^2)$ according to $N(0, 1)$. In short, $P(x_S, Z_1 = z_1, Z_2 = z_2) = P(x_S|Z_1 = z_1, Z_2 = z_2)P(Z_1 = z_1, Z_2 = z_2)$,[1] where

$$\begin{aligned} &P(x_S|Z_1 = z_1, Z_2 = z_2) \\ &\quad = P(x_{S^1}|Z_1 = z_1)P(x_{S^2}|Z_2 = z_2)P(x_{S\setminus(S^1\cup S^2)}) \\ &\quad = \prod_{s\in S^1} G(x_s; z_1, 1) \prod_{s\in S^2} G(x_s; z_2, 1) \\ &\qquad \times \prod_{s\in S\setminus(S^1\cup S^2)} G(x_s; 0, 1) \end{aligned} \tag{1}$$

and $G(x; \mu, \sigma)$ stands for the normal probability density (mean μ and standard deviation σ) evaluated at x.

Imagine now that we are presented with a sample image generated by the model. Our problem is to decide whether or not the image contains the letter L. We will devise and analyze several decision rules, and later relate the conclusions to more general and relevant models, and to the discussion of clutter, context, and computation.

Optimal Decision Rule In this example, the presence of an L is equivalent to the presence of horizontal and vertical bars, i.e. the event $\{Z_1 = 1\} \cap \{Z_2 = 1\}$. This suggests thresholding on the posterior probability, $\mathcal{S}_G(x_S) \doteq P(Z_1 = 1, Z_2 = 1|X_S = x_S)$:

Declare "L" if $\mathcal{S}_G(x_S) > t$ and "not L" if $\mathcal{S}_G(x_S) \leq t$.

The threshold governs the tradeoff between false alarms and missed detections, and the set of all thresholds defines the ROC curve. The decision rule is optimal in that it minimizes the probability of missed detections at any given probability of false alarms. (This follows from the Neyman-Pearson Lemma and the observation that $\mathcal{S}_G(x_S)$ is a monotone increasing function of the likelihood ratio $\frac{P(x_S|L\ present)}{P(x_S|L\ not\ present)}$.)

Observations:

1.

$$\begin{aligned} &\mathcal{S}_G(x_S) \\ &\quad = \frac{P(x_S|Z_1 = 1, Z_2 = 1)P(Z_1 = 1, Z_2 = 1)}{\sum_{z_1=0}^{1}\sum_{z_2=0}^{1} P(x_S|Z_1 = z_1, Z_2 = z_2)P(Z_1 = z_1, Z_2 = z_2)} \\ &\quad = \frac{P(x_{S^1}|Z_1 = 1)P(x_{S^2}|Z_2 = 1)P(Z_1 = 1, Z_2 = 1)}{\sum_{z_1=0}^{1}\sum_{z_2=0}^{1} P(x_{S^1}|Z_1 = z_1)P(x_{S^2}|Z_2 = z_2)P(Z_1 = z_1, Z_2 = z_2)} \end{aligned} \tag{2}$$

which follows from Bayes' formula and the decomposition in (1).

2. Also by (1):

$$\begin{aligned} \mathcal{S}_G(x_S) &= P(Z_1 = 1, Z_2 = 1|X_S = x_S) \\ &= P(Z_1 = 1|X_S = x_S) \\ &\quad \times P(Z_2 = 1|Z_1 = 1, X_S = x_S) \\ &= P(Z_1 = 1|X_{S^1} = x_{S^1}, X_{S^2} = x_{S^2}) \\ &\quad \times P(Z_2 = 1|Z_1 = 1, X_{S^2} = x_{S^2}) \end{aligned} \tag{3}$$

As this is the product of two conditional probabilities, it suggests a sequential version of the test $\mathcal{S}_G(x_S) > t$. In particular, if $P(Z_1 = 1|X_{S^1} = x_{S^1}, X_{S^2} = x_{S^2}) > t$ fails then there is no point in computing $P(Z_2 = 1|Z_1 = 1, X_{S^2} = x_{S^2})$, since $\mathcal{S}_G(x_S) \leq P(Z_1 = 1|X_{S^1} = x_{S^1}, X_{S^2} = x_{S^2})$. If it does not fail, then we compute $P(Z_2 = 1|Z_1 = 1, X_{S^2} = x_{S^2})$ and compare the product $P(Z_1 = 1|X_{S^1} = x_{S^1}, X_{S^2} = x_{S^2})P(Z_2 = 1|Z_1 = 1, X_{S^2} = x_{S^2})$ to t. We will return to this shortly.

Template Matching The problem with $\mathcal{S}_G$ is that it can not be computed, at least not in general, as is evident from examining equation (2). The denominator is the full likelihood, meaning a mixture over *every possible explanation* of the data. The mixture has one term for "{L}∩$\{X_S = x_S\}$" and all the rest for "{not an L}∩$\{X_S = x_S\}$." It is one thing to compute (or estimate) a reasonable likelihood for the singleton events "L" or "nothing there," and quite another to compute a likelihood for the compound event "not an L."

[1] Pixel data, x_s, $s \in S$, could be modeled as continuous or discrete. Unless we are conditioning, we will avoid writing $X_s = x_s$, so that for example $P(x_S|Z_1 = z_1, Z_2 = z_2)$ could be the evaluation of a density or a probability mass function.

A sensible, and in one way or another much-used, alternative is to approximate "{not an L}$\cap\{X_S = x_S\}$" by "{nothing there}$\cap\{X_S = x_S\}$," i.e. to use the statistic

$$\mathcal{S}_T(x_S) \doteq \frac{P(x_S|Z_1=1, Z_2=1)P(Z_1=1, Z_2=1)}{\sum_{z=0}^{1} P(x_S|Z_1=z, Z_2=z)P(Z_1=z, Z_2=z)} \tag{4}$$

Observations:

1. By the same reasoning used for $\mathcal{S}_G$:

$$\mathcal{S}_T(x_S) = \frac{P(x_{S^1}|Z_1=1)P(x_{S^2}|Z_2=1)P(Z_1=1, Z_2=1)}{\sum_{z=0}^{1} P(x_{S^1}|Z_1=z)P(x_{S^2}|Z_2=z)P(Z_1=z, Z_2=z)} \tag{5}$$

2. Thresholding on $\mathcal{S}_T$ is the same as thresholding on the likelihood ratio

$$\frac{P(x_{S^1}|Z_1=1)P(x_{S^2}|Z_2=1)}{P(x_{S^1}|Z_1=0)P(x_{S^2}|Z_2=0)} \tag{6}$$

which might be more familiar.

3. $\mathcal{S}_T$ is optimal under a different probability, $\tilde{P}$, on the latent variables:

$$\mathcal{S}_T(x_S) = \tilde{P}(Z_1=1, Z_2=1|X_S=x_S) \tag{7}$$

where $\tilde{P}(x_S, Z_1=z_1, Z_2=z_2)$ is

$$P(x_S, Z_1=z_1, Z_2=z_2|(Z_1, Z_2)=(1,1) \text{ or } (Z_1, Z_2)=(0,0)) \tag{8}$$

i.e. pretending that the world has only two states, "object" or "nothing."

Testing for Parts This is a modification of the factored representation, (3), of the optimal decision rule. The second term, $P(Z_2=1|Z_1=1, X_{S^2}=x_{S^2})$, is local to S^2. On the other hand, the first term, $P(Z_1=1|X_{S^1}=x_{S^1}, X_{S^2}=x_{S^2})$, is global in the sense that it involves the evaluation of data likelihoods for every assignment of states to the remaining parts. There is only one remaining part in this example, but in general the number of assignments of states will be exponential in the number of parts. These observations suggest a third statistic, derived by approximating $P(Z_1=1|X_{S^1}=x_{S^1}, X_{S^2}=x_{S^2})$ with the corresponding local probability $P(Z_1=1|X_{S^1}=x_{S^1})$:

$$\mathcal{S}_P(x_S) \doteq P(Z_1=1|X_{S^1}=x_{S^1}) \times P(Z_2=1|Z_1=1, X_{S^2}=x_{S^2}) \tag{9}$$

The test $\mathcal{S}_P(x_S) > t$ can be performed sequentially. The first test is for the first part ($P(Z_1=1|X_{S^1}=x_{S^1}) > t$), *ignoring* information in the pixel data about the second part. If $P(Z_1=1|X_{S^1}=x_{S^1}) > t$ then the second part is tested (via $P(Z_1=1|X_{S^1}=x_{S^1})P(Z_2=1|Z_1=1, X_{S^2}=x_{S^2}) > t$), using the pixels in the support of the second part and a probability that is computed in the context of the presumed presence of the first part.

Foveal Limit We want to compare these three strategies. The optimal serves as a benchmark against which the performance of template matching and parts-based testing can be measured. The setup is simple enough that both mathematical and exhaustive computational analyses are possible. Concerning mathematical analysis, we will examine relative performances by comparing the ROC curves in the limit as the density of pixels goes to infinity (the "foveal limit").[2] In other words, spacing between pixels of the uniform grid S in Fig. 3 is decreased to zero.

All three approaches are perfect in the foveal limit. Hence the areas under the three ROC curves converge to one. We will compare the rates at which the areas *above* the ROC curves converge to zero. Obviously, neither template matching nor parts-based testing can do better than optimal. But which of the two suboptimal approaches should we expect to better approximate optimal performance? One way to think about this is to anticipate the primary sources of confusions for each of the suboptimal tests. Consider two sets of circumstances. In the first, both parts are present ($Z_1=1$ and $Z_2=1$) but one or the other of the parts is substantially degraded. A template takes into account all of the data, and from this point of view the situation is not really different from a uniform, but less severe, degradation of the entire L. As for parts-based testing, it is vulnerable to missing the degraded part, especially when the degraded part is tested first.[3] On the other hand either part can appear alone, and in such cases template matching, in that it is essentially making a forced decision between $Z_1=Z_2=1$ and $Z_1=Z_2=0$, is vulnerable to false alarms.

It turns out that the consequences of the second circumstance dominate, overwhelmingly.

To make this precise, given a statistic $\mathcal{S}=\mathcal{S}(x_S)$, let $\mathcal{A}_{\mathcal{S}}$ be the area above the ROC curve generated by the test $\mathcal{S}(x_S) > t$. Necessarily, $\mathcal{A}_{\mathcal{S}_T} \geq \mathcal{A}_{\mathcal{S}_G}$ and $\mathcal{A}_{\mathcal{S}_P} \geq \mathcal{A}_{\mathcal{S}_G}$ (Neyman Pearson). Concerning the simple two-part world constructed above:

[2] We are using "foveal," and later "saliency," descriptively rather than to suggest a biological model of saccades and fixations.

[3] The reader might be tempted to conclude that the optimal test should suffer the same vulnerability to a degraded part, in light of the sequential representation of (3). But in contrast to the parts-based test, the first test in the sequential version of the optimal decision function takes into account the appearances of *both parts*.

Theorem *If $P(Z_1 = z_1, Z_2 = z_2) > 0$ for every pair $(z_1, z_2) \in \{0, 1\}^2$, then in the foveal limit*

1. $\mathcal{A}_{\mathcal{S}_P}/\mathcal{A}_{\mathcal{S}_G}$ *remains bounded*;
2. $\mathcal{A}_{\mathcal{S}_T}/\mathcal{A}_{\mathcal{S}_P} \to \infty$ *exponentially fast.*

Remarks

1. The conclusions are the same if

$$\mathcal{S}_P(x_S) = P(Z_1 = 1|X_{S^1} = x_{S^1}) \times P(Z_2 = 1|Z_1 = 1, X_{S^2} = x_{S^2}) \quad (10)$$

is replaced by

$$\mathcal{S}'_P(x_S) \doteq P(Z_1 = 1|X_{S^1} = x_{S^1})P(Z_2 = 1|X_{S^2} = x_{S^2}) \quad (11)$$

i.e. if the discovery of the first part is ignored in testing for the second part. The theorem is about parts and background, and not context per se. If parts are real, i.e. themselves reusable and not just a proxy for an object, and if the evidence is strong, then parts-based testing is nearly optimal (in the sense that $\mathcal{A}_{\mathcal{S}_P}/\mathcal{A}_{\mathcal{S}_G}$ is bounded) and much better than templates. Context comes in when we use (10) instead of (11) and, later (see "saliency"), a modification of (10) for better finite-resolution performance. The experimental evidence strongly favors (10) over (11), and saliency-based testing over (10), but a proof might require very different methods.
2. The conclusions are also the same if the last term in (1), $P(x_{S\backslash(S^1\cup S^2)})$, is replaced by an arbitrary conditional distribution, $P(x_{S\backslash(S^1\cup S^2)}|X_{S^1\cup S^2} = x_{S^1\cup S^2})$. But then it is much harder to explore extensions, as we shall do shortly in Sect. 3.
3. Concerning extensions, there is nothing particularly special about the Gaussian distribution. The proof is constructed for more general distributions, as laid out at the beginning of Appendix A.

Proof Generically, for any statistic $\mathcal{S}(x_S)$

$$\mathcal{A}_{\mathcal{S}} = Prob\{\mathcal{S}(X_S) < \mathcal{S}(\tilde{X}_S)\} \quad (12)$$

where X_S and $\tilde{X}_S$ are independent samples from $P(x_S|\{Z_1 = 1\} \cap \{Z_2 = 1\})$ and $P(x_S|\{\{Z_1 = 1\} \cap \{Z_2 = 1\}\}^C)$, respectively. This, along with the various independence assumptions and a standard one-dimensional large-deviation result, makes the comparisons relatively straightforward. The detailed proof is in the Appendix A. □

The three ROC curves can be computed numerically. Figure 4 explores performance of all three classifiers as a function of resolution, for small and moderate pixel densities as well as the larger densities corresponding to the "foveal limit" covered by the theorem. At the lowest density there are only two pixels in the support of the horizontal bar and four in the support of the vertical. Template matching is not far from optimal, and better than parts-based testing. The order is already reversed when there are just four and eight pixels representing the horizontal and vertical bars, respectively. With eight and sixteen pixels, parts-based testing is nearly indistinguishable from optimal, and substantially outperforms the template model. A glance at higher resolutions confirms that the template model converges to perfect classification much more slowly than the other two.

Saliency Testing $P(Z_1 = 1|X_{S^1} = x_{S^1})P(Z_2 = 1|Z_1 = 1, X_{S^2} = x_{S^2})$ is not the same as testing $P(Z_2 = 1|X_{S^2} = x_{S^2})P(Z_1 = 1|Z_2 = 1, X_{S^1} = x_{S^1})$. In $\mathcal{S}_P(x_S) \doteq P(Z_1 = 1|X_{S^1} = x_{S^1})P(Z_2 = 1|Z_1 = 1, X_{S^2} = x_{S^2})$, the second part has an advantage: if Z_1 and Z_2 are the parts of an object, then typically $P(Z_2 = 1|Z_1 = 1) > P(Z_2 = 1)$. Therefore the test $\mathcal{S}_P(x_S) > t$ is more vulnerable to a degraded view of the first part then the second. With these observations in mind, it might make sense to look first at the part for which the evidence is strongest. (Occlusion, for which we would advocate a layered or fully 3-D model, is another matter.) When there are many parts, instead of just two, then the idea can be applied recursively: first test for the most salient part, then test for the conditionally most salient part, given the part already found, and so on. The result is order dependent because tests for all but the first part are conditioned on the presence of the earlier parts. Here we take a closer look at these ideas, by extending the theorem from two parts to an arbitrary number of parts, and the factorization from a fixed-order to a data-dependent order. We illustrate with some additional experiments.

Suppose that our object, call it $\mathcal{O}$, is made of N parts rather than just two. Extending the notation in the obvious way, we let $Z_k \in \{0, 1\}$ indicate the absence or presence of the kth part, $1 \le k \le N$, let $S^k \subseteq S$ be the pixels in the support of the kth part, and let X_{S^k} be the corresponding pixel intensities. We assume that $S^k \cap S^l = \emptyset$, for all $k \ne l$, though there will be some discussion of this later. To ease the notation, we will write v_1^k for a vector $(v_1, \dots, v_k)$, and use shorthand like $v_1^k = 1$ for $v_i = 1$, $\forall 1 \le i \le k$. The joint distribution of $X_S, Z_1, Z_2, \dots, Z_N$ (or X_S, Z_1^N for short) is modeled by extension of the L model: $P(x_S, Z_1^N = z_1^N) = P(x_S|Z_1^N = z_1^N)P(Z_1^N = z_1^N)$, where

$$\begin{aligned} P(x_S|Z_1^N = z_1^N) &= P(x_{S\backslash\bigcup_{k=1}^n S^k}) \prod_{k=1}^N P(x_{S^k}|Z_k = z_k) \\ &= G(x_{S\backslash\bigcup_{k=1}^n S^k}; 0, 1) \prod_{k=1}^N G(x_{S^k}; z_k, 1) \end{aligned} \quad (13)$$

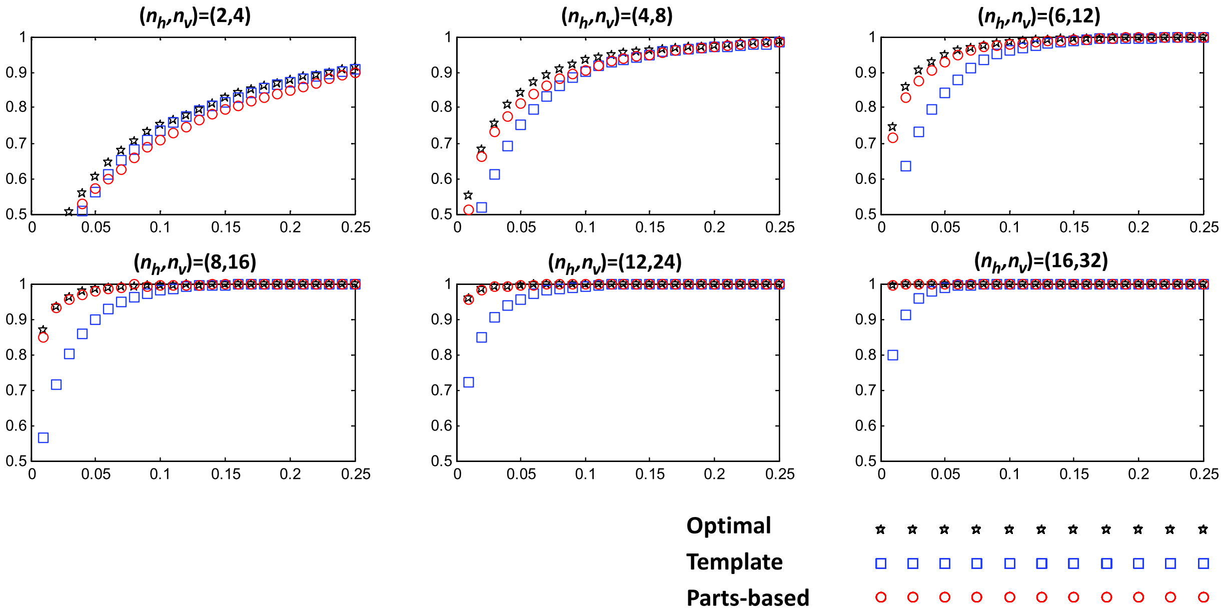


Fig. 4 (Color online) **Illustration of Comparison Theorem.** Each panel contains three ROC curves, for the optimal ($\mathcal{S}_G(x_S) > t$, *black stars*), for template matching ($\mathcal{S}_T(x_S) > t$, *blue squares*) and for parts-based testing ($\mathcal{S}_P(x_S) > t$, *red circles*). Resolution is progressively increased, *left-to-right* and *top-to-bottom* ("foveal limit"). In each panel the numbers of pixels on the *horizontal* and *vertical bars* (the "supports") are indicated by (n_h, n_v) (so $n_h = |S^1|$ and $n_v = |S^2|$). At low resolution, $(n_h, n_v) = (2, 4)$, template matching outperforms parts-based testing. At higher resolutions parts-based testing is better, and nearly optimal. Template matching is slow to converge to perfect performance

and $G(x_A; \mu, \sigma)$ stands for $\prod_{s \in A} G(x_s; \mu, \sigma)$ (iid normal) for any $A \subseteq S$. Finally, we say that the object $\mathcal{O}$ is present if and only if all of its parts are present.[4]

The extensions of the optimal decision rule ($\mathcal{S}_G(x_S) > t$) and template matching ($\mathcal{S}_T(x_S) > t$) involve straightforward changes in the statistics:

$$\mathcal{S}_G(x_S) \doteq P(Z_1^N = 1 | X_S = x_S) \tag{14}$$

and

$$\mathcal{S}_T(x_S) \doteq \frac{P(x_S|Z_1^N = 1)P(Z_1^N = 1)}{P(x_S|Z_1^N = 0)P(Z_1^N = 0) + P(x_S|Z_1^N = 1)P(Z_1^N = 1)} \tag{15}$$

All of the various observations about these two statistics, made earlier for the case $N = 2$, still hold when $N \geq 2$, with obvious changes in the formulas.

As for parts-based testing ($\mathcal{S}_P(x_S) > t$), we want to make a more fundamental change by extending it to allow for a data-dependent factorization. The factors can be thought of as defining a sequence of tests, one for each part. The first test is directed at the "most salient part," by which we mean the most probable part when only local evidence is taken into account (i.e. based on X_{S^k} and not X_S):

$$k_1 \doteq \arg\max_k P(Z_k = 1 | X_{S^k} = x_{S^k}) \tag{16}$$

The first test is $P(Z_{k_1} = 1 | X_{S^{k_1}} = x_{S^{k_1}}) > t$. If it succeeds, then we compute the most salient of the remaining parts, but now in the context of $Z_{k_1} = 1$:

$$k_2 \doteq \arg\max_{k \neq k_1} P(Z_k = 1 | Z_{k_1} = 1, X_{S^k} = x_{S^k}) \tag{17}$$

The second test is $P(Z_{k_1} = 1 | X_{S^{k_1}} = x_{S^{k_1}}) P(Z_{k_2} = 1 | Z_{k_1} = 1, X_{S^{k_2}} = x_{S^{k_2}}) > t$. Iterating through N parts generates a *random sequence* $k_i = k_i(X_S)$, $i = 1, 2, \ldots, N$

$$k_i \doteq \arg \max_{k \notin \{k_1, \ldots, k_{i-1}\}} P(Z_k = 1 | Z_{k_1}^{k_{i-1}} = 1, X_{S^k} = x_{S^k}) \tag{18}$$

and defines a random (data-dependent) factorization, and a new statistic $\mathcal{S}_Q$:

$$\mathcal{S}_Q(x_S) \doteq \prod_{i=1}^{N} P(Z_{k_i} = 1 | Z_{k_1}^{k_{i-1}} = 1, X_{S^{k_i}} = x_{S^{k_i}}) \tag{19}$$

[4] Among other things, (13) says that X_{S^k} is sufficient for z_k, for each $k = 1, 2, \ldots, N$. But *this is different* from saying that given X_{S^k} the state of Z_k can be inferred equally well by ignoring the rest of the data, $X_{S \setminus S^k}$. It can not. There is a difference between ordinary sufficiency and "Bayesian sufficiency." What is true is that $X_{\bigcup_{k=1}^N S^k}$ is Bayesian sufficient for $(Z_1, \ldots, Z_N)$.

where we stretch the notation, somewhat, by letting $Z_{k_1}^{k_{i-1}}$ stand for $(Z_{k_1}, Z_{k_2}, \dots, Z_{k_{i-1}})$.

Concerning asymptotics, the same theoretical result holds for $\mathcal{S}_Q$ as for $\mathcal{S}_P$:

Corollary 1 *If* $P(Z_1^N = z_1^N) > 0$ *for every* $z_1^N \in \{0,1\}^N$, *then in the foveal limit*

1. $\mathcal{A}_{\mathcal{S}_Q}/\mathcal{A}_{\mathcal{S}_G}$ *remains bounded*;
2. $\mathcal{A}_{\mathcal{S}_T}/\mathcal{A}_{\mathcal{S}_Q} \to \infty$ *exponentially fast.*

There is very little different from the proof as already given in the Appendix A. We forgo the details.

Concerning finite-resolution performance, the difference between a fixed-order parts-based test and a test ordered by saliency can be explored by returning to the simple L world and performing the same experiment as reported in Fig. 4, but with

$$\mathcal{S}_Q(x_S) = P(Z_{k_1} = 1 | X_{S^{k_1}} = x_{S^{k_1}}) \times P(Z_{k_2} = 1 | Z_{k_1} = 1, X_{S^{k_2}} = x_{S^{k_2}}) \tag{20}$$

instead of

$$\mathcal{S}_P(x_S) = P(Z_1 = 1 | X_{S^1} = x_{S^1}) \times P(Z_2 = 1 | Z_1 = 1, X_{S^2} = x_{S^2}) \tag{21}$$

Figure 5 is identical to Fig. 4, except that the random-order parts-based test ($\mathcal{S}_Q(x_S) > t$) was used. Comparing to Fig. 4, parts-based testing is now nearly equivalent to optimal testing at all resolutions. It is intuitive that visiting parts in the order of saliency is better than using a fixed order, especially in the low-resolution domain, and no doubt something can be proven along these lines. But the approach will need to be different, since the analysis behind the theorem and corollary is asymptotic, in the foveal (high-resolution) limit.

As an additional illustration, we chose a problem that is still easy enough that versions of the optimal classifier and template matching classifier can be computed, but is no longer entirely artificial. Starting with an ASCII (e-book) version of Ernest Hemingway's novel "For Whom the Bell Tools," we built an image of every page by choosing a resolution (pixel dimensions per page) and creating a JPEG image. The first page, at an intermediate resolution, can be seen on the left-hand side of Fig. 6. There is no noise, per se, but the moderate resolution and random positioning of characters relative to pixels creates significant degradation. The task was to search the manuscript for specific words, "at" and "the" in the experiments reported in the figure.

For each character in a word we built an appearance model by assuming (wrongly) that the pixels in the support are iid, with different distributions for the two conditions "character present" and "character absent". Every page was partitioned into blocks, within which there could be a character, a symbol, or a blank. For each letter in the word and each of the two conditions, "present" or absent", the manuscript was used to build two empirical distributions for the pixels in the character's support. These empirical distributions were used for the data model. Notice that typically other characters would be present when a given character was absent—the iid assumption is crude. Referring to Fig. 6, then, the "optimal decision rule" isn't really optimal since the data model is merely an approximation.

These approximations do not seem to have affected the relative performances, as compared to the L example in which the model was exact. ROC performance of parts testing with saliency-based ordering is indistinguishable from the (approximate) optimal, and substantially better than template matching, for detecting "at" and "the" (right-hand side of the figure). Obviously, there are many levels of relevant context, including word pairs that are more or less usual, sentence structure, the topic of a paragraph or a chapter, and even an author's style and preferred vocabulary. For that matter, the characters themselves create contextual effects as seen at the level of strokes. We could just as well have used strokes, instead of characters, as the basic parts.

We end this chapter with several observations about $\mathcal{S}_Q(x_S)$, related to computation and interpretation:

1. **Computation.** How much does it cost to use the factorization in equation (19)? The first step already requires examining all of the data in the support of object $\mathcal{O}$, i.e. each of x_{S^k}, $k = 1, 2, \dots, N$ in order to calculate each of $P(Z_k = 1 | X_{S^k} = x_{S^k})$. This was avoided in the fixed-order scheme. On the other hand, once these N conditional probabilities have been computed the remaining tests come down to computing the conditional probability of one part of $\mathcal{O}$ given the presence of a set of other parts of $\mathcal{O}$, $P(Z_{k_i} = 1 | Z_{k_1}^{k_{i-1}} = 1)$. (See below.) This is the contextual term, involving a computation on the prior distribution but not the data. It may be cheap or expensive, but in any case some version of this computation, either in closed form or by approximation (e.g. belief propagation), is unavoidable in any formulation of contextual reasoning.

   The connection between the calculation of

   $$P(Z_{k_i} = 1 | Z_{k_1}^{k_{i-1}} = 1, X_{S^{k_i}} = x_{S^{k_i}}) \tag{22}$$

   and the calculation of

   $$P(Z_{k_i} = 1 | X_{S^{k_i}} = x_{S^{k_i}}) \tag{23}$$

   is through the likelihood ratio

   $$l = \frac{P(x_{S^{k_i}} | Z^{k_i} = 1)}{P(x_{S^{k_i}} | Z^{k_i} = 0)} \tag{24}$$

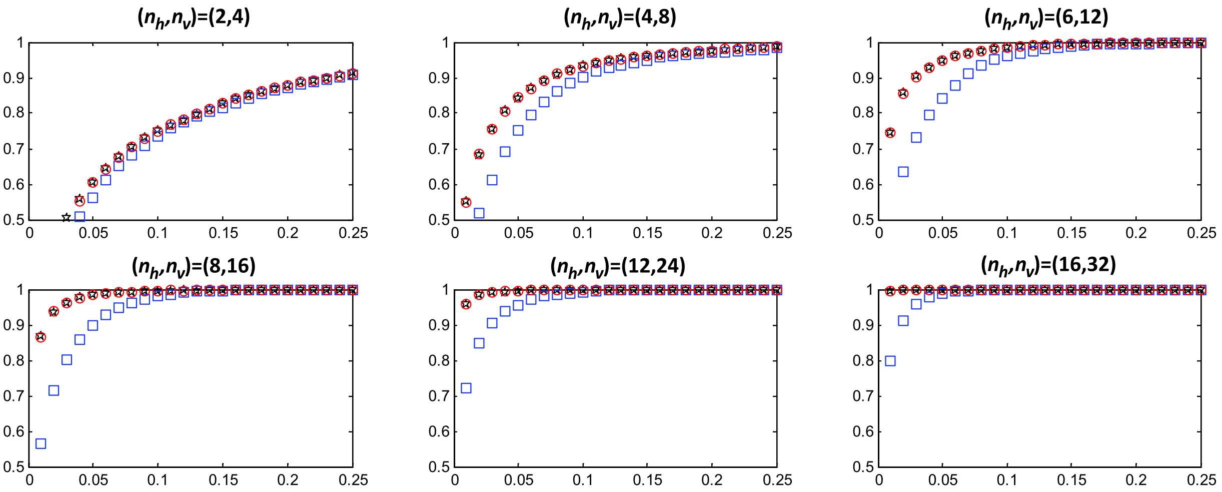
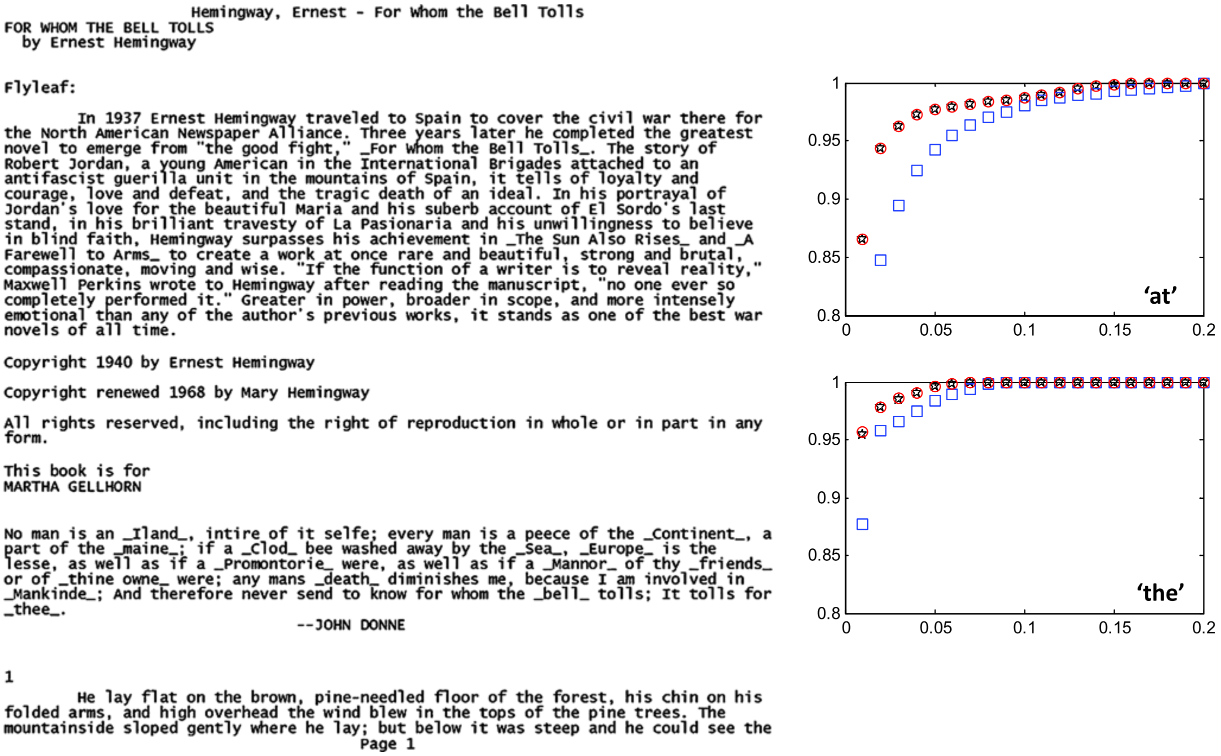


Fig. 5 **Saliency.** Same as Fig. 4, except that parts are tested in the order of saliency (i.e. their conditional probabilities given only *local* pixel data) for the parts-based decision function. Compared to Fig. 4, parts-based testing is now essentially optimal at all resolutions. (Color scheme same as in Fig. 4)



Hemingway, Ernest - For Whom the Bell Tolls

FOR WHOM THE BELL TOLLS
by Ernest Hemingway

Flyleaf:

In 1937 Ernest Hemingway traveled to Spain to cover the civil war there for the North American Newspaper Alliance. Three years later he completed the greatest novel to emerge from "the good fight," _For Whom the Bell Tolls_. The story of Robert Jordan, a young American in the International Brigades attached to an antifascist guerilla unit in the mountains of Spain, it tells of loyalty and courage, love and defeat, and the tragic death of an ideal. In his portrayal of Jordan's love for the beautiful Maria and his suberb account of El Sordo's last stand, in his brilliant travesty of La Pasionaria and his unwillingness to believe in blind faith, Hemingway surpasses his achievement in _The Sun Also Rises_ and _A Farewell to Arms_ to create a work at once rare and beautiful, strong and brutal, compassionate, moving and wise. "If the function of a writer is to reveal reality," Maxwell Perkins wrote to Hemingway after reading the manuscript, "no one ever so completely performed it." Greater in power, broader in scope, and more intensely emotional than any of the author's previous works, it stands as one of the best war novels of all time.

Copyright 1940 by Ernest Hemingway

Copyright renewed 1968 by Mary Hemingway

All rights reserved, including the right of reproduction in whole or in part in any form.

This book is for
MARTHA GELLHORN

No man is an _Iland_, intire of it selfe; every man is a peece of the _Continent_, a part of the _maine_; if a _Clod_ bee washed away by the _Sea_, _Europe_ is the lesse, as well as if a _Promontorie_ were, as well as if a _Mannor_ of thy _friends_ or of _thine owne_ were; any mans _death_ diminishes me, because I am involved in _Mankinde_; And therefore never send to know for whom the _bell_ tolls; It tolls for _thee_.
--JOHN DONNE

1

He lay flat on the brown, pine-needled floor of the forest, his chin on his folded arms, and high overhead the wind blew in the tops of the pine trees. The mountainside sloped gently where he lay; but below it was steep and he could see the
Page 1

Fig. 6 **Word Search.** On the left is an image of the first page of Ernest Hemingway's novel, "For Whom the Bell Tolls." The ASCII e-book was converted to a relatively low-resolution JPEG image. The image was used to search for all instances of the words "at" and "the" in the novel. A simple model was estimated and the ROC performance of each of the three decision algorithms (optimal, template, and parts-based) was computed. (Here, "optimal" means relative to the estimated model.) Testing of parts, ordered by saliency, was indistinguishable from the optimal test, and substantially better than template matching for both word searches. (Color scheme same as in Fig. 4)

The pixel data enters only through l, and in fact both $P(Z_{k_i}=1|Z_{k_1}^{k_{i-1}}=1, X_{S^{k_i}}=x_{S^{k_i}})$ and $P(Z_{k_i}=1|X_{S^{k_i}}=x_{S^{k_i}})$ are simple monotone functions of l. Specifically, if

$$M(q,l) \doteq \frac{lq}{lq+(1-q)} \tag{25}$$

then $P(Z_{k_i}=1|X_{S^{k_i}}=x_{S^{k_i}})=M(q,l)$ with $q=P(Z_{k_i}=1)$ and $P(Z_{k_i}=1|Z_{k_1}^{k_{i-1}}=1, X_{S^{k_i}}=x_{S^{k_i}})=M(q,l)$ with $q=P(Z_{k_i}=1|Z_{k_1}^{k_{i-1}}=1)$.

In other words, $P(Z_{k_i}=1|Z_{k_1}^{k_{i-1}}=1, X_{S^{k_i}}=x_{S^{k_i}})$ is just $P(Z_{k_i}=1|X_{S^{k_i}}=x_{S^{k_i}})$, but with $P(Z_{k_i}=1)$ "tilted" by context into $P(Z_{k_i}=1|Z_{k_1}^{k_{i-1}}=1)$.

2. **Sequential Testing.** There is a clear advantage to testing $\mathcal{S}_P(x_S)>t$ one part at a time. Unless the object is ubiquitous, one of the tests, $P(Z_1=1|X_{S^1}=x_{S^1})\cdots P(Z_k=1|Z_1^{k-1}=1, X_{S^k}=x_{S^k})>t$, will typically fail well before $k=N$. Most of the computation is avoided. By contrast, even if $\mathcal{S}_Q(x_S)>t$ fails early not much is gained, since deciding which part to test first already costs order N operations. (The conclusion is quite different when the parts, Z_k, have more than just two states, as we shall see in Sect. 3.)
3. **Context and Departure from Independence.** Typically, $P(Z_{k_i}=1|Z_{k_1}^{k_{i-1}}=1)>P(Z_{k_i}=1)$, by virtue of the accumulating evidence for the object $\mathcal{O}$, and since $M(q,l)$ is also monotone in q, the threshold for object k_i is effectively reduced if it comes late in the testing. This then is the condition for a contextual effect: $P(Z_{k_i}=1|Z_{k_1}^{k_{i-1}}=1)>P(Z_{k_i}=1)$, which is reminiscent of Barlow's (1994) and other discussions of learning in hierarchies (Harrison 2005; Zhu and Mumford 2006; Ommer and Buhmann 2007; Chen et al. 2007). Iterating the expression, and dropping the cumbersome ordering (which is irrelevant to the interpretation of the inequality) we arrive at the condition

$$\frac{P(Z_1=1, Z_2=1,\ldots,Z_N=1)}{\prod_{k=1}^N P(Z_k=1)}>1 \tag{26}$$

which implies an analogous expression for any subset of the parts of $\mathcal{O}$. The ratio on the left-hand side of (26) is a measure of departure from independence, in the direction of a strong positive contextual effect. (What Barlow calls a suspicious coincidence.) As a first cut to developing a learning rule for hierarchical systems it is not unreasonable to take the empirical estimate of this ratio as evidence in favor of explicitly representing the composition of these parts, and thereby leading to the "discovery" of the object $\mathcal{O}$. From this viewpoint, (unsupervised) learning might be thought of as a continuing effort to remove otherwise unexplained dependencies through the introduction of new compositions.
4. **Local Likelihoods.** It is instructive to compare the optimal statistic to the parts-based statistic. Unlike the parts-based strategy, where each ordering of the visitation to parts defines a different statistic, the statistic defining the optimal strategy (i.e. $\mathcal{S}_G$) is independent of ordering, whether or not the ordering is random. In particular

$$\mathcal{S}_G(x_S)=\prod_{i=1}^N P(Z_{k_i}=1|Z_{k_1}^{k_{i-1}}=1, X_{S^{k_i}}=x_{S^{k_i}},\ldots, X_{S^{k_N}}=x_{S^{k_N}}) \tag{27}$$

for any permutation k of the indices $\{1,2,\ldots,N\}$, which follows by straightforward extension of the reasoning used to derive (3). Interpreted sequentially, every test of a part looks ahead to the pixel data of the remaining, untested parts. By comparison, $\mathcal{S}_Q$ ((19) and $\mathcal{S}_P$ (same, but with fixed order) are local at every stage of the computation. Contextual influence from the pixel data associated with parts not yet visited is ignored, relying only on the contextual influence from the parts already visited and presumed present. It is not hard to see, then, that the re-ordering of the part-visitation schedule according to saliency can have a substantial impact on performance, consistent with the experiments.

3 Generalizations

Objects are not usually represented as single conjunctions of rigid parts, as in the thought experiment studied in Sect. 2. Almost by definition, invariance calls for a *disjunction* of conjunctions, e.g. representing variation in structure (three strokes in a European-style seven or two in an American-style seven), pose (a large or small seven in this or that vicinity), appearance (one color or another, matt or glossy surface) and so on. Hierarchical models use "and/or graphs" (Zhu and Mumford 2006; Chen et al. 2007), or equivalently composition or production rules (Shieber 1992; Jin and Geman 2006; Felzenszwalb and McAllester 2010), to build layers of disjunctions of conjunctions. Invariance emerges progressively with each layer, as the number of instantiations increases exponentially. Here we will take a first step in the direction of invariance by considering an object made of parts with multiple states, each part, for example, with its own pose.

A given collection of parts (such as a vertical and a horizontal stroke or two image patches that look like left and right eyes) does not generally guarantee the presence of an object. The arrangement of poses, for example, can be coincidental. Usually there are cues in the relationships among the parts, possibly in their relational coordinates or in the matching of certain attributes, that give evidence one way or another. We will discuss generalizations in which the presence of parts does not, in and of itself, imply the presence of an object.

Consider a simple two-layer hierarchy consisting of an object $\mathcal{O}$ and its N parts, $Z_1, \ldots, Z_N$. $\mathcal{O}$ is binary: absent ($\mathcal{O}=0$) or present ($\mathcal{O}=1$). To bring in the idea of multiple instantiations, we expand the range of the latent variables $Z_1, \ldots, Z_N$ from binary ($Z_k \in \{0,1\}$, part k is absent or present) to n-ary ($Z_k \in \{0,1,\ldots,n\}$, part k is absent or present and in state $Z_k > 0$). The different states of a part might represent a partitioning of pose space and/or a selection of styles or rendering models. In the illustration below, they represent the positions and sizes of the front and back tires in a broadside view of a car. Of course the parts need not have the same numbers of states ($n = n^k$, $k = 1, 2, \ldots, N$), but to keep the notation as simple as possible we will assume that they do.

In general $P(\mathcal{O}=1|Z_1^N > 0) \neq 1$, in keeping with the idea that the presence of the parts is not enough to guarantee the presence of the object. On the other hand, we will assume that the object is always represented by these same N parts, so that $P(Z_1^N = z_1^N|\mathcal{O}=1)$ concentrates on $z_1^N > 0$. Disjunctions of conjunctions of different collections of parts will be briefly discussed in Sect. 4.

Staying as close as possible to our earlier notation, we let $S^k_{z_k}$ be the support of part k when $Z_k = z_k > 0$. (For example, different poses of the same part will typically have different supports.) When $z_k = 0$, we interpret $S^k_{z_k}$ as the empty set. In order to reformulate the theorem, without an undo list of conditions and details, we make a simplifying assumption: the support of one part never intersects the support of another, i.e. for any $k \neq l$

$$S_i^k \cap S_j^l = \emptyset \quad \forall i, j \in \{1, 2, \ldots, n\} \tag{28}$$

Of course eventually we will have to contend with multiple objects in arbitrary poses, and then assumptions of this sort are untenable. But the message, concerning the ROC performance of parts-based testing, will likely turn out to be the same.

If we let $S^k = \bigcup_{i=1}^n S_i^k$ then $S^k \cap S^l = \emptyset$ whenever $k \neq l$, and we can still think of S^k as the support of part k. The generalization of (1), the conditional data model, is

$$\begin{aligned} P(x_S|Z_1^N = z_1^N) &= \left(\prod_{k=1}^N P(x_{S^k}|Z_k = z_k)\right) P(x_{S\setminus\bigcup_{k=1}^N S^k}) \\ &= \left(\prod_{k=1}^N \prod_{s\in S^k_{z_k}} p_1(x_s)\right) \prod_{s\in S\setminus\bigcup_{k=1}^N S^k_{z_k}} p_o(x_s) \end{aligned} \tag{29}$$

for some "target" and "null" distributions p_1 and p_o (both formerly Gaussian). What then are the analogs of the optimal, the saliency-ordered parts-based, and the template-based decision rules? What, if anything, can be said about their relative merits, along the lines of the theorem?

The optimal approach doesn't really change: threshold on $P(\mathcal{O}=1|X_S = x_s)$, which, in light of (29), is the same as thresholding on the statistic

$$\mathcal{S}_G(x_S) = \frac{\sum_{z_1^N} \prod_{k=1}^N \prod_{s\in S^k_{z_k}} \frac{p_1(x_s)}{p_o(x_s)} P(Z_1^N = z_1^N|\mathcal{O}=1) P(\mathcal{O}=1)}{\sum_{o=0}^1 \sum_{z_1^N} \prod_{k=1}^N \prod_{s\in S^k_{z_k}} \frac{p_1(x_s)}{p_o(x_s)} P(Z_1^N = z_1^N|\mathcal{O}=o) P(\mathcal{O}=o)} \tag{30}$$

keeping in mind that $P(Z_1^N = z_1^N|\mathcal{O}=1) = 0$ unless $z_k > 0$ at every k, and the convention $S_0^k = \emptyset$.

As in Sect. 2, we interpret template matching as an optimal decision rule under the all-or-none approximation of P, i.e. either $\mathcal{O}=1$, or when $\mathcal{O}=0$ no parts are present ($Z_1 = Z_2 = \cdots = Z_N = 0$). Accordingly, by modification of the expression for the optimal statistic $\mathcal{S}_G(x_S)$:

$$\mathcal{S}_T(x_S) = \frac{\sum_{z_1^N} \prod_{k=1}^N \prod_{s\in S^k_{z_k}} \frac{p_1(x_s)}{p_o(x_s)} P(Z_1^N = z_1^N|\mathcal{O}=1) P(\mathcal{O}=1)}{P(Z_1^N = 0|\mathcal{O}=0) P(\mathcal{O}=0) + \sum_{z_1^N} \prod_{k=1}^N \prod_{s\in S^k_{z_k}} \frac{p_1(x_s)}{p_o(x_s)} P(Z_1^N = z_1^N|\mathcal{O}=1) P(\mathcal{O}=1)} \tag{31}$$

To motivate the parts-based test, first re-write the optimal test, $\mathcal{S}_G(x_S) > t$, in a sequential form:

$$\begin{aligned}\mathcal{S}_G(x_S) &= P(Z_1 > 0|X_S = x_S) \cdot P(Z_2 > 0|Z_1 > 0, X_S = x_S)\\ &\quad \cdots P(Z_N > 0|Z_1^{N-1} > 0, X_S = x_S)\\ &\quad \cdot P(\mathcal{O} = 1|Z_1^N > 0, X_S = x_S) > t \end{aligned} \tag{32}$$

In words, test for parts sequentially, and then test the arrangement of parts for the object. Since the accumulating product is non-increasing, we quit if and when it drops to or below the threshold t. In detail, the ith test is

$$\begin{aligned}&\sum_{z_1>0} P(Z_1 = z_1|X_S = x_S)\\ &\quad \cdot \sum_{z_2>0} P(Z_2 = z_2|Z_1 = z_1, X_S = x_S)\\ &\quad \cdots \sum_{z_i>0} P(Z_i = z_i|Z_1^{i-1} = z_1^{i-1}, X_S = x_S) > t\end{aligned} \tag{33}$$

If the Nth test passes, then the final test is for $\mathcal{O}$:

$$\begin{aligned}&\sum_{z_1>0} P(Z_1 = z_1|X_S = x_S)\\ &\quad \cdot \sum_{z_2>0} P(Z_2 = z_2|Z_1 = z_1, X_S = x_S)\\ &\quad \cdots \sum_{z_N>0} P(Z_N = z_N|Z_1^{N-1} = z_1^{N-1}, X_S = x_S)\\ &\quad \cdot P(\mathcal{O} = 1|Z_1^N = z_1^N) > t\end{aligned} \tag{34}$$

The parts-based version is similar, but always "local." Replace $\sum_{z_i>0} P(Z_i = z_i|Z_1^{i-1} = z_1^{i-1}, X_S = x_S)$ in the optimal test (34) by $\sum_{z_i>0} P(Z_i = z_i|Z_1^{i-1} = z_1^{i-1}, X_{S^i} = x_{S^i})$:

$$\begin{aligned}\mathcal{S}_P(x_S) &= \sum_{z_1>0} P(Z_1 = z_1|X_{S^1} = x_{S^1})\\ &\quad \cdot \sum_{z_2>0} P(Z_2 = z_2|Z_1 = z_1, X_{S^2} = x_{S^2})\\ &\quad \cdots \sum_{z_N>0} P(Z_N = z_N|Z_1^{N-1} = z_1^{N-1}, X_{S^N} = x_{S^N})\\ &\quad \cdot P(\mathcal{O} = 1|Z_1^N = z_1^N) > t\end{aligned} \tag{35}$$

As in the binary case ($Z_k \in \{0, 1\}$), context enters, one step at a time, through the tilted probabilities $P(Z_i = z_i|Z_1^{i-1} = z_1^{i-1})$.

As for saliency, mimic (16) and (18): the first part to test, k_1, is the one with highest local conditional probability,

$$k_1 \doteq \arg\max_k \sum_{z_k>0} P(Z_k = z_k|X_{S^k} = x_{S^k}) \tag{36}$$

Thereafter, test the remaining part with highest local conditional probability, but now operating in the context of the confirmed parts, i.e. under the tilted distribution:

$$\begin{aligned}k_i &\doteq \arg\max_{k\notin\{k_1,\ldots,k_{i-1}\}} \sum_{z_{k_1}>0} P(Z_{k_1} = z_{k_1}|X_{S^{k_1}} = x_{S^{k_1}})\\ &\quad \cdot \sum_{z_{k_2}>0} P(Z_{k_2} = z_{k_2}|Z_{k_1} = z_{k_1}, X_{S^{k_2}} = x_{S^{k_2}})\\ &\quad \cdots \sum_{z_k>0} P(Z_k = z_k|Z_{k_1}^{k_{i-1}} = z_{k_1}^{k_{i-1}}, X_{S^k} = x_{S^k})\end{aligned} \tag{37}$$

Which brings us, finally, to the statistic

$$\begin{aligned}&\mathcal{S}_Q(x_S)\\ &\quad = \sum_{z_{k_1}>0} P(Z_{k_1} = z_{k_1}|X_{S^{k_1}} = x_{S^{k_1}})\\ &\quad\quad \cdot \sum_{z_{k_2}>0} P(Z_{k_2} = z_{k_2}|Z_{k_1} = z_{k_1}, X_{S^{k_2}} = x_{S^{k_2}})\\ &\quad\quad \cdots \sum_{z_{k_N}>0} P(Z_{k_N} = z_{k_N}|Z_{k_1}^{k_{N-1}} = z_{k_1}^{k_{N-1}}, X_{S^{k_N}} = x_{S^{k_N}})\\ &\quad\quad \cdot P(\mathcal{O} = 1|Z_1^N = z_1^N)\end{aligned} \tag{38}$$

There are two issues at hand: multiple states of the parts (e.g. representing multiple poses), and the possibility that a coincidental arrangement of the parts might be indistinguishable from the presence of the object. We address these, separately, in the following paragraphs, and then conclude this section with a discussion of computation and the coarse-to-fine (sequential) implementation of parts-based testing.

When the Parts Guarantee the Object In this case $P(\mathcal{O} = 1|Z_1 > 0, \ldots, Z_N > 0) = 1$ and the theorem is essentially unchanged:

Corollary 2 *If* $P(Z_1^N = z_1^N) > 0$ *for every* $z_1^N \in \{0, 1, \ldots, n\}^N$, *and if* $P(\mathcal{O} = 1|Z_1 > 0, \ldots, Z_N > 0) = 1$, *then in the foveal limit*

1. $\mathcal{A}_{\mathcal{S}_Q}/\mathcal{A}_{\mathcal{S}_G}$ *remains bounded*;
2. $\mathcal{A}_{\mathcal{S}_T}/\mathcal{A}_{\mathcal{S}_Q} \to \infty$ *exponentially fast.*

Remark We focus on $\mathcal{S}_Q$, but from the point of view of asymptotics, at least up to the limitations of our analysis, $\mathcal{S}_P$ and $\mathcal{S}_Q$ (and other part-based methods–see remarks following the theorem) are indistinguishable.

Coincidences If $P(\mathcal{O} = 1|Z_1 > 0, \ldots, Z_N > 0) = 1$, then detecting $\mathcal{O}$ amounts to detecting each of the N parts. What happens if $P(\mathcal{O} = 1|Z_1 > 0, \ldots, Z_N > 0) < 1$? Then there will be configurations, $z_1^N > 0$, for which $P(\mathcal{O} = 0|Z_1^N =$

$z_1^N) > 0$ (allowing for coincidences). If for these configurations it is also the case that $P(\mathcal{O}=1|Z_1^N=z_1^N)>0$, then $Z_1^N=z_1^N$ is ambiguous. Perfect performance is impossible, even with perfect knowledge of the parts. In and of themselves, two tires of the right size and separation certainly do not guarantee a car. Maybe they are on separate cars, or not on any cars at all. The joint probability of $(\mathcal{O}, Z_1^N)$ defines its own optimal ROC curve, through the decision function "declare $\mathcal{O}=1$ if $P(\mathcal{O}|Z_1^N=z_1^N)>t$, and $\mathcal{O}=0$ otherwise." This is the ROC curve based on a perfect knowledge of the parts, which is exactly what we have in the foveal limit. In short, the optimal decision function given the pixel data, $\mathcal{S}_G(x_S)>t$, approaches the optimal decision function $P(\mathcal{O}|Z_1^N=z_1^N)>t$ in the foveal limit.

Let $\mathcal{S}_{G'}(z_1^N)=P(\mathcal{O}|Z_1^N=z_1^N)$ and let $\mathcal{A}_{\mathcal{S}_{G'}}$ be the area above the corresponding ROC curve. If $P(\mathcal{O}=1|Z_1>0,\dots,Z_N>0)=1$ then $\mathcal{A}_{\mathcal{S}_{G'}}=0$, but in general $\mathcal{A}_{\mathcal{S}_{G'}}>0$. In order to compare $\mathcal{A}_{\mathcal{S}_G}$, $\mathcal{A}_{\mathcal{S}_Q}$, and $\mathcal{A}_{\mathcal{S}_T}$ in the general case, we compare $\mathcal{D}_{\mathcal{S}_G}\doteq\mathcal{A}_{\mathcal{S}_G}-\mathcal{A}_{\mathcal{S}_{G'}}$, $\mathcal{D}_{\mathcal{S}_Q}\doteq\mathcal{A}_{\mathcal{S}_Q}-\mathcal{A}_{\mathcal{S}_{G'}}$, and $\mathcal{D}_{\mathcal{S}_T}\doteq\mathcal{A}_{\mathcal{S}_T}-\mathcal{A}_{\mathcal{S}_{G'}}$, all of which are necessarily nonnegative and all of which go to zero in the foveal limit:

Corollary 3 *If* $P(Z_1^N=z_1^N)>0$ *for every* $z_1^N\in\{0,1,\dots,n\}^N$, *then in the foveal limit*

1. $\mathcal{D}_{\mathcal{S}_Q}/\mathcal{D}_{\mathcal{S}_G}$ *remains bounded*;
2. $\mathcal{D}_{\mathcal{S}_T}/\mathcal{D}_{\mathcal{S}_Q}\to\infty$ *exponentially fast.*

The proofs are, again, more-or-less straightforward extensions of the special case treated in Appendix A.

Illustration We built a primitive car detector. We collected 163 side views of cars, 122 from Caltech 101 and an additional 42 from Google Images, and 200 natural images, used as negative examples. Cars were modeled as consisting of nothing more than two tires, one in the front and one in the back. Hence there were only two parts, Z_1 and Z_2, which coded the poses of the front and back tires, respectively. Each tire could be at one of five scales and at any position within the image. Consequently, the number of states of each part was about five times the number of pixels in the image. The joint distribution, $P(Z_1=z_1, Z_2=z_2|\mathcal{O}=1)$, on tire placements given the presence of a (sideways) car ("$\mathcal{O}=1$") was restricted to concentrate on pairs (z_1,z_2) for which z_1 and z_2 represented the same scale, and to depend only on the scale-corrected position of one tire with respect to the other. Hence $P(\cdot|\mathcal{O}=1)$ was scale and location invariant. In the absence of a car, Z_1 and Z_2 were assumed to be independent, $P(Z_1=z_1,Z_2=z_2|\mathcal{O}=0)=P(Z_1=z_1|\mathcal{O}=0)P(Z_2=z_2|\mathcal{O}=0)$.

For each of the three decision rules, the data-dependent calculations come down to likelihood ratios,

$$\frac{P(x_{S^k_{z_k}}|Z_k=z_k)}{P(x_{S^k_{z_k}}|Z_k=0)} \tag{39}$$

for part k ($k\in\{1,2\}$) in pose z_k ($z_k\in\{1,\dots,n\}$, $n\approx 5\times\#\text{pixels}$). We modeled these ratios directly via a sufficient statistic, namely the normalized correlation between a template and the data. Since our interest is in comparative performance, rather than good performance, per se, we used coarse, 8 × 8, templates (T_1 and T_2), one for each tire. The likelihood ratio (39) of the data $x_{S^k_{z_k}}$ becomes the likelihood ratio of the correlation, $C(T_k, x_{S^k_{z_k}})$, which is well approximated as a backwards exponential divided by a zero-mean normal:

$$\frac{P(x_{S^k_{z_k}}|Z_k=z_k)}{P(x_{S^k_{z_k}}|Z_k=0)}=\frac{\lambda e^{-\lambda(1-C(T_k,x_{S^k_{z_k}}))}}{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-C(T_k,x_{S^k_{z_k}})^2/2\sigma^2}} \tag{40}$$

with MLE estimates for λ and σ.

The problem is small enough and the model simple enough that the entire integration (double summation) could be computed for each of the three approaches, resulting in the three ROC curves shown in Fig. 7. Obviously, as with the word search in the previous example, there is nothing truly optimal about the "optimal" decision rule, except to say that it would be best possible if the model were actually true. Still, the relative merits of the three approaches, as captured by the three curves, are consistent with the more controlled experiments reported in Figs. 4, 5, and 6. Even under an incorrect model, it is better to evaluate the data globally ($\mathcal{S}_G$) than locally ($\mathcal{S}_Q$), and a properly configured local test is better than an all-or-nothing template ($\mathcal{S}_T$). What is more, local tests ordered by saliency perform nearly as well as global tests.

Remarks on Computation In the special case discussed in Sect. 2, $Z_k\in\{0,1\}$, $k=1,2,\dots,N$, all of these statistics, $\mathcal{S}_T$, $\mathcal{S}_P$, and $\mathcal{S}_Q$, are computable in the sense that they involve computation that is linear in N. Only $\mathcal{S}_G$, which includes in the denominator a summation over $z_1^N\in\{0,1\}^N$ (fully accounting for background), is intractable (exponential in N), barring restrictive assumptions. When $Z_k\in\{0,1,\dots,n\}$, $n>1$, all four statistics involve exponential sums, over the positive set $z_1^N\in\{1,2,\dots,n\}^N$ in the cases of $\mathcal{S}_T$, $\mathcal{S}_P$, and $\mathcal{S}_Q$, and the full set $z_1^N\in\{0,1,\dots,n\}^N$ in the case of $\mathcal{S}_G$.

But $\mathcal{S}_P$, and $\mathcal{S}_Q$ offer some relief: both can be tested against t ($\mathcal{S}_P>t$, $\mathcal{S}_Q>t$) through a sequence of tests, each test against the same threshold t, for subsets of i parts,

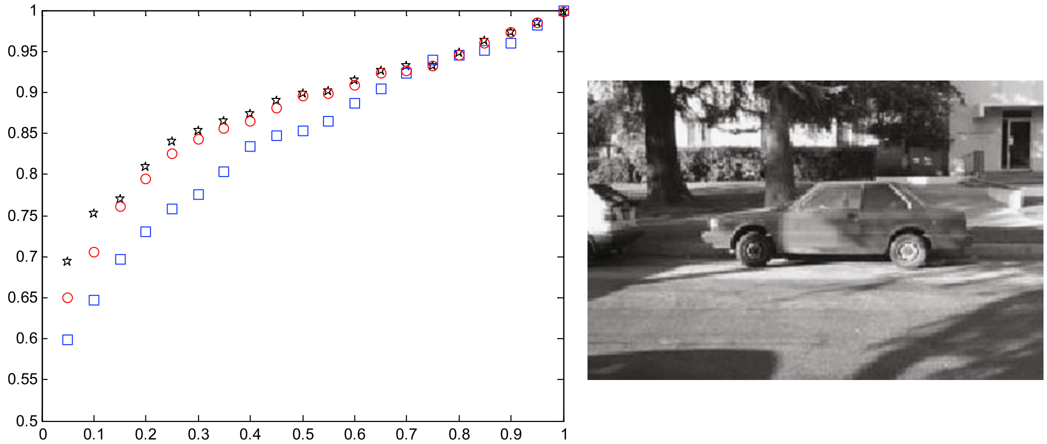


Fig. 7 **Tires and Cars.** We constructed a simple car detector based on a search for suitably positioned pairs of tires. The probability model includes crude appearance models, one for each tire, and the empirical distribution on the relative coordinates between the tires. ROC curves are shown for the three decision rules: optimal (with respect to the estimated model), template, and parts-based testing ordered by saliency. Parts-based testing performs nearly as well as the (approximated) optimal test. (Color scheme same as in Fig. 4)

$i = 1, 2, \ldots, N$. Explicitly, the ith test for $\mathcal{S}_P$ is

$$\sum_{z_1>0} P(Z_1 = z_1 | X_{S^1} = x_{S^1}) \cdot \sum_{z_2>0} P(Z_2 = z_2 | Z_1 = z_1, X_{S^2} = x_{S^2}) \cdots \sum_{z_i>0} P(Z_i = z_i | Z_1^{i-1} = z_1^{i-1}, X_{S^i} = x_{S^i}) > t \qquad (41)$$

and for $\mathcal{S}_Q$ is

$$\sum_{z_{k_1}>0} P(Z_{k_1} = z_{k_1} | X_{S^{k_1}} = x_{S^{k_1}}) \cdot \sum_{z_{k_2}>0} P(Z_{k_2} = z_{k_2} | Z_{k_1} = z_{k_1}, X_{S^{k_2}} = x_{S^{k_2}}) \cdots \sum_{z_{k_i}>0} P(Z_{k_i} = z_{k_i} | Z_{k_1}^{k_{i-1}} = z_{k_1}^{k_{i-1}}, X_{S^{k_i}} = x_{S^{k_i}}) > t \qquad (42)$$

In each case, the ith test costs some multiple of n^i operations.[5] We can expect that in most applications the first test, or at least an early test, will fail most of the time. On these occasions the computational advantage over $\mathcal{S}_T$ and $\mathcal{S}_G$ will be exponential.

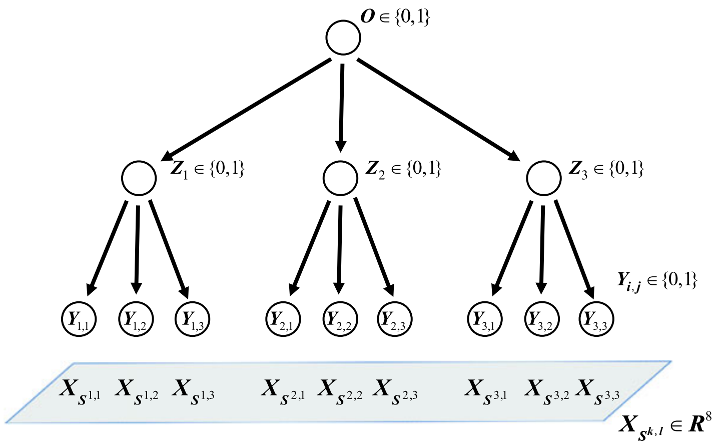
But $\mathcal{S}_P$, and $\mathcal{S}_Q$ won't always fail early, and every detection of $\mathcal{O}$ encounters $O(n^N)$ operations. Some form of pruning is needed, and in fact can be quite effective, but this is of course model dependent. If, for example, the states $z_1, z_2, \ldots, z_N$ represent pose, then $P(Z_1^N|\mathcal{O} = 1)$ amounts to a shape model, which will typically restrict the effective number of terms in (41)and (42) for all but the smallest i. As more parts pass, and i increases, the uncertainty in pose will decrease and the effective number of terms will be a negligible fraction of n^i.

4 Hierarchical Models

Hierarchical models are well-suited for exploiting multilevel contextual constraints. To what extent can the parts-based tests studied in Sects. 2 and 3 be reformulated for more general compositional systems? Here we will explore one such reformulation, that amounts to a recursive application of the localization and tilting operations already studied in Sects. 2 and 3.

In keeping with our earlier discussions, we will experiment with a model sufficiently simple that the optimal detection algorithm can be computed and used as a benchmark. Figure 8 depicts a nearly minimal hierarchical model. In addition to the pixel data, there are three levels of latent variables, representing reusable parts that presumably could be found either by themselves or as components of other parts, objects, or object compositions. The single top-level variable, $\mathcal{O}$, is assumed to indicate the presence or absence, with $\mathcal{O} = 1$ or $\mathcal{O} = 0$, of an object of interest. All computations are feasible, thanks to the Markov structure and small problem size. As already discussed, the object would typically be modeled as a hierarchy of disjunctions of conjunctions, via an and/or graph or a related architecture, to accommodate

[5] The cost of choosing the ith most salient part, in $\mathcal{S}_Q$, multiplies the cost by some additional factor, but not bigger than N.



$\boldsymbol{O} \in \{0,1\}$
$\boldsymbol{Z}_1 \in \{0,1\}$
$\boldsymbol{Z}_2 \in \{0,1\}$
$\boldsymbol{Z}_3 \in \{0,1\}$
$\boldsymbol{Y}_{i,j} \in \{0,1\}$
$Y_{1,1}$ $Y_{1,2}$ $Y_{1,3}$ $Y_{2,1}$ $Y_{2,2}$ $Y_{2,3}$ $Y_{3,1}$ $Y_{3,2}$ $Y_{3,3}$
$X_{S^{1,1}}$ $X_{S^{1,2}}$ $X_{S^{1,3}}$ $X_{S^{2,1}}$ $X_{S^{2,2}}$ $X_{S^{2,3}}$ $X_{S^{3,1}}$ $X_{S^{3,2}}$ $X_{S^{3,3}}$
$X_{S^{k,l}} \in \boldsymbol{R}^8$

Fig. 8 Model of an Instantiation of Object $\mathcal{O}$. In hierarchical models, objects are usually represented as layered disjunctions of conjunctions. A particular instantiation is a collection of conjunctions, such as the one shown here used to illustrate parts-based computation in hierarchical models. The probability model, with its simple, directed and acyclic graphical structure, admits a computationally feasible calculation of an optimal statistic for testing $\mathcal{O}$, and provides a benchmark for evaluating strategies based on testing parts and testing templates

multiple presentations for invariant recognition. We will focus on verifying a single instantiation of $\mathcal{O}$, which is to say a selected hierarchy of conjunctions, such as the one shown in the figure.

The model, a Bayesian net, is specified through its conditional probabilities. Specifically, we assume that $P(Z_1^3 = z_1^3|\mathcal{O} = 1)$ is one if $z_1 = z_2 = z_3 = 1$, and zero otherwise. On the other hand, if $\mathcal{O} = 0$ then the parts are independent:

$$\begin{aligned} P(Z_1^3 = z_1^3|\mathcal{O} = 0) \\ &= P(Z_1 = z_1|\mathcal{O} = 0)P(Z_2 = z_2|\mathcal{O} = 0) \\ &\quad \times P(Z_3 = z_3|\mathcal{O} = 0) \end{aligned} \tag{43}$$

Similarly, for each $i = 1, 2, 3$, $P(Y_{i,1} = Y_{i,2} = Y_{i,3} = 1|Z_i = 1) = 1$, and

$$\begin{aligned} P(Y_{i,1} = y_{i,1}, Y_{i,2} = y_{i,2}, Y_{i,3} = y_{i,3}|Z_i = 0) \\ &= P(Y_{i,1} = y_{i,1}|Z_i = 0)P(Y_{i,2} = y_{i,2}|Z_i = 0) \\ &\quad \times P(Y_{i,3} = y_{i,3}|Z_i = 0) \end{aligned} \tag{44}$$

The distribution on the latent variables is then fully specified by the a prior probability that $\mathcal{O} = 1$, which we take to be 0.01, the conditional probability that $Z_i = 1$ given that $\mathcal{O} = 0$, which we take to be 0.1 for every $1 \le i \le 3$, and the conditional probability that $Y_{i,j} = 1$ given that $Z_i = 0$, which we take to be 0.2 for every $1 \le i, j \le 3$.

The full generative model is completed by specifying a conditional data distribution, for which we follow the previous examples and use

$$\begin{aligned} P(x_S|Y_{i,j} = y_{i,j}, 1 \le i, j \le 3, Z_i = z_i, 1 \le i \le 3, \mathcal{O} = o) \\ &= P(x_S|Y_{i,j} = y_{i,j}, 1 \le i, j \le 3) \\ &= \prod_{1\le i,j\le 3}\prod_{s\in S^{i,j}} G(x_s; y_{i,j}, 1) \prod_{s\in S\setminus\bigcup_{1\le i,j\le 3} S^{i,j}} G(x_s; 0, 1) \end{aligned} \tag{45}$$

Since $P(\mathcal{O} = 1|Z_1^3 = 1) < 1$, perfect performance is impossible, even in the foveal limit—see Sect. 3.

The computation of $\mathcal{S}_G(x_S) = P(\mathcal{O} = 1|X_S = x_s)$ is a standard exercise in graphical models. Concerning template matching, recall that $\mathcal{S}_T$ is $\tilde{P}(\mathcal{O} = 1|X_S = x_s)$, where $\tilde{P}$ is the conditional distribution under P given that all of the latent variables are either on (i.e. 1) or off (i.e. 0):

$$\begin{aligned} \mathcal{S}_T(x_S) &= P(\mathcal{O} = 1)P(Z. = 1|\mathcal{O} = 1)P(Y_{.,.} = 1|Z. = 1) \\ &\quad \times \prod_{1\le i,j\le 3}\prod_{s\in S^{i,j}} G(x_s; 1, 1) \\ &\quad /\Big\{P(\mathcal{O} = 0)P(Z. = 0|\mathcal{O} = 0)P(Y_{.,.} = 0|Z. = 0) \\ &\quad \times \prod_{1\le i,j\le 3}\prod_{s\in S^{i,j}} G(x_s; 0, 1) \\ &\quad + P(\mathcal{O} = 1)P(Z. = 1|\mathcal{O} = 1)P(Y_{.,.} = 1|Z. = 1) \\ &\quad \times \prod_{1\le i,j\le 3}\prod_{s\in S^{i,j}} G(x_s; 1, 1)\Big\} \end{aligned} \tag{46}$$

where $Z.$ is the same as Z_1^3, $Y_{i,.}$ means $(Y_{i,1}, Y_{i,2}, Y_{i,3})$, and $Y_{.,.}$ means $(Y_{1,1}, Y_{1,2}, \ldots, Y_{3,3})$. Using $l(v)$ to denote the likelihood ratio $G(v; 1, 1)/G(v; 0, 1)$, and putting in the specific parameters of the model:

$$\mathcal{S}_T(x_S) = \frac{(.01)\prod_{1\le i,j\le 3}\prod_{s\in S^{i,j}} l(x_s)}{(.01)(.1)^3(.2)^9 + (.01)\prod_{1\le i,j\le 3}\prod_{s\in S^{i,j}} l(x_s)} \tag{47}$$

As for $\mathcal{S}_P$, this is a localized version of a suitable factorization of $\mathcal{S}_G$, as in the earlier examples. Start with the top composition in Fig. 8, through which Z_1, Z_2, and Z_3 compose to make $\mathcal{O}$:

$$\begin{aligned}\mathcal{S}_G(x_S) &= P(\mathcal{O}=1|X_S=x_S)\\ &= P(\mathcal{O}=1|X_{S^{\cdot,\cdot}}=x_{S^{\cdot,\cdot}})\\ &= P(\mathcal{O}=1|Z_{\cdot}=1)P(Z_{\cdot}=1|X_{S^{\cdot,\cdot}}=x_{S^{\cdot,\cdot}})\\ &= P(\mathcal{O}=1|Z_{\cdot}=1)P(Z_1=1|X_{S^{\cdot,\cdot}}=x_{S^{\cdot,\cdot}})\\ &\quad\cdot P(Z_2=1|Z_1=1, X_{S^{\cdot,\cdot}}=x_{S^{\cdot,\cdot}})\\ &\quad\cdot P(Z_3=1|Z_1=1, Z_2=1, X_{S^{\cdot,\cdot}}=x_{S^{\cdot,\cdot}}) \qquad (48)\end{aligned}$$

where we use the analogous "dot" notation for sets: $S^{i,\cdot}=S^{i,1}\cup S^{i,2}\cup S^{i,3}$, and $S^{\cdot,\cdot}=\bigcup_{1\le i,j\le 3}S^{i,j}$. The top-level factorization of $\mathcal{S}_T$ is obtained by replacing the exact factorization of $\mathcal{S}_G$ by its local approximation. Observing that $X_{S^{i,\cdot}}$ is the local data for Z_i (see Fig. 8):

$$\begin{aligned}\mathcal{S}_G(x_S) \longrightarrow\ & P(\mathcal{O}=1|Z_{\cdot}=1)P(Z_1=1|X_{S^{1,\cdot}}=x_{S^{1,\cdot}})\\ &\cdot P(Z_2=1|Z_1=1, X_{S^{2,\cdot}}=x_{S^{2,\cdot}})\\ &\cdot P(Z_3=1|Z_1=1, Z_2=1, X_{S^{3,\cdot}}=x_{S^{3,\cdot}}) \quad (49)\end{aligned}$$

The process is recursive. Each of the three factors $P(Z_i=1|Z_1^{i-1}=1, X_{S^{i,\cdot}}=x_{S^{i,\cdot}})$, $1\le i\le 3$, is of the type we started with, $P(\mathcal{O}=1|X_S=x_S)$, in that they are conditional probabilities of compositions given the data associated with their component parts. The only difference is that the prior probability of the composition, $P(Z_i=1)$, is tilted by "earlier" tests and the participation of Z_i in the composition of $\mathcal{O}$, $P(Z_i=1)\to P(Z_i=1|Z_1^{i-1}=1)$. Apply parts-based factorization to $P(Z_i=1|Z_1^{i-1}=1, X_{S^{i,\cdot}}=x_{S^{i,\cdot}})$:

$$\begin{aligned}P(Z_i=1|Z_1^{i-1}&=1, X_{S^{i,\cdot}}=x_{S^{i,\cdot}})\\ &= P(Z_i=1|Z_1^{i-1}=1, Y_{i,\cdot}=1)\\ &\quad\cdot P(Y_{i,\cdot}=1|Z_1^{i-1}=1, X_{S^{i,\cdot}}=x_{S^{i,\cdot}})\\ &= P(Z_i=1|Z_1^{i-1}=1, Y_{i,\cdot}=1)\\ &\quad\cdot P(Y_{i,1}=1|Z_1^{i-1}=1, X_{S^{i,\cdot}}=x_{S^{i,\cdot}})\\ &\quad\cdot P(Y_{i,2}=1|Z_1^{i-1}=1, Y_{i,1}=1, X_{S^{i,\cdot}}=x_{S^{i,\cdot}})\\ &\quad\cdot P(Y_{i,3}=1|Z_1^{i-1}=1, Y_{i,1}^{i,2}=1, X_{S^{i,\cdot}}=x_{S^{i,\cdot}}) \qquad (50)\end{aligned}$$

And then localize:

$$\begin{aligned}P(Z_i=1|Z_1^{i-1}&=1, X_{S^{i,\cdot}}=x_{S^{i,\cdot}})\\ \longrightarrow\ & P(Z_i=1|Z_1^{i-1}=1, Y_{i,\cdot}=1)\\ &\cdot P(Y_{i,1}=1|Z_1^{i-1}=1, X_{S^{i,1}}=x_{S^{i,1}})\\ &\cdot P(Y_{i,2}=1|Z_1^{i-1}=1, Y_{i,1}=1, X_{S^{i,2}}=x_{S^{i,2}})\\ &\cdot P(Y_{i,3}=1|Z_1^{i-1}=1, Y_{i,1}^{i,2}=1, X_{S^{i,3}}=x_{S^{i,3}}) \qquad (51)\end{aligned}$$

There are no more levels and the terms in (51) can now be evaluated directly.

The factorization can be made more transparent by exposing its recursive nature, for which we now introduce some additional notation. This will also serve to more easily explain the randomly ordered (saliency) version of the statistic (i.e. $\mathcal{S}_Q(x_S)$, see below). Recall that $l(v)=G(v;1,1)/G(v;0,1)$. We will use $l_{i,j}$ as shorthand for the likelihood ratio of the data in the support of $Y_{i,j}$:

$$l_{i,j}=\prod_{s\in S^{i,j}} l(x_s) \qquad (52)$$

Next, we introduce a general form for the (fixed-order) parts-based factorization of a composition, say 'W': $\mathcal{S}_P^W(x_{S^W};q)$, where W is an object or part, x_{S^W} is the data in the support of W, and q is the a priori or context-tilted probability of W. In terms of these representations, and using the particular parameters of the current model, the factorization defined by (49) and (51) can be summarized as follows:

$$\mathcal{S}_P(x_S)=\mathcal{S}_P^{\mathcal{O}}(x_{S^{\cdot,\cdot}};P(\mathcal{O}=1)) \qquad (53)$$

where

$$\begin{aligned}&\mathcal{S}_P^{\mathcal{O}}(x_{S^{\cdot,\cdot}};\alpha)\\ &\quad=\frac{\alpha}{\alpha+(1-\alpha)(.1)^3}\prod_{i=1}^{3}\mathcal{S}_P^{Z_i}\left(x_{S^{i,\cdot}};\frac{\alpha+(1-\alpha)(.1)^i}{\alpha+(1-\alpha)(.1)^{i-1}}\right)\end{aligned} \qquad (54)$$

$$\begin{aligned}&\mathcal{S}_P^{Z_i}(x_{S^{i,\cdot}};\beta)\\ &\quad=\frac{\beta}{\beta+(1-\beta)(.2)^3}\prod_{j=1}^{3}\mathcal{S}_P^{Y_{i,j}}\left(x_{S^{i,j}};\frac{\beta+(1-\beta)(.2)^j}{\beta+(1-\beta)(.2)^{j-1}}\right)\end{aligned} \qquad (55)$$

and finally

$$\mathcal{S}_P^{Y_{i,j}}(x_{S^{i,j}};\gamma)=\frac{\gamma l_{i,j}}{\gamma l_{i,j}+(1-\gamma)} \qquad (56)$$

If $\mathcal{O}$, itself, were a part in a composition, then, depending on the order of factors, $\mathcal{S}_P^{\mathcal{O}}(x_{S^{\cdot,\cdot}};\alpha)$ would be evaluated at the tilted probability of $\mathcal{O}=1$ rather than its a priori probability, $\alpha=P(\mathcal{O}=1)$. Every time an evaluation is finished, of $\mathcal{S}_P^{Y_{i,j}}(x_{S^{i,j}};\gamma)$, $\mathcal{S}_P^{Z_i}(x_{S^{i,\cdot}};\beta)$, or $\mathcal{S}_P^{\mathcal{O}}(x_{S^{\cdot,\cdot}};\alpha)$, the existing product of terms can be tested against the common threshold t, as described in previous sections. Generally speaking, an early exit from the recursion will be the rule rather than the exception.

$\mathcal{S}_Q$ has the same form as $\mathcal{S}_P$, but $i \in \{1,2,3\}$ is replaced by a permutation k_i, and $j \in \{1,2,3\}$ is replaced by permutations $m_{i,j}$, one for each i. There is therefore one permutation for each of the four compositions in the model. As in previous examples, the orderings of indices are determined by a measure of saliency, applied in a greedy (and now recursive) fashion.

First, for each part Z_i the composition of subparts $Y_{i,\cdot}$ is evaluated, locally, and ordered by saliency, under the a priori probability of $Z_i = 1$. The most likely of the three parts, factored in the order used for the saliency-based evaluation of subparts, defines the first set of terms in $\mathcal{S}_Q$. This fixes k_1 (the first part) and $m_{k_1,1}$, $m_{k_1,2}$, and $m_{k_1,3}$ (the order of its subparts). The process is repeated for the remaining two parts, but now under the tilted probability $P(Z_i = 1|Z_{k_1} = 1)$, for each remaining i. This fixes k_2, and $m_{k_2,1}$, $m_{k_2,2}$, and $m_{k_2,3}$. A third pass assigns k_3 to the remaining part, and $m_{k_3,1}$, $m_{k_3,2}$, and $m_{k_3,3}$ to the ordered sequence of subparts under the distribution tilted by $P(Z_{k_3} = 1|Z_{k_1}^{k_2} = 1)$. Pseudo code can be found in Appendix B.

Figure 9 shows the results of detection experiments using the statistics $\mathcal{S}_G$, $\mathcal{S}_T$, $\mathcal{S}_P$, and $\mathcal{S}_Q$. There were eight pixels in the support of each subpart ($|S^{k,l}| = 8$). The top panel compares $\mathcal{S}_G$, $\mathcal{S}_T$, and $\mathcal{S}_P$, and the bottom compares $\mathcal{S}_G$, $\mathcal{S}_T$, and $\mathcal{S}_Q$. The trend is the same as in the earlier experiments. Parts testing beats template matching, and saliency improves on parts testing.

Both of the parts-based procedures (fixed order and random order) have suggestive connections to so-called bottom-up and top-down computing (see Amit and Geman 1998; Borenstein and Ullman 2002; Kokkinos et al. 2006; Wu and Zhu 2010, where very similar connections are made). Overall, the computation is bottom-up, in that subparts are tested before parts, and parts are tested before testing $\mathcal{O}$. Along the way, the test of a subpart changes the probability of its parent part (bottom-up), which in turn changes the probabilities of sibling subparts (top-down). A change in the probability of a part changes the probability of $\mathcal{O}$ (bottom-up), which in turn changes the probabilities of other parts and their subparts (top-down). More layers and more elaborate models provide more avenues for both types of computing. Thus, for example, a given part might be the child of many parents (reusability), and a given parent might be instantiated by any one of multiple children sets (as in and/or graphs and other mechanisms for modeling invariance). In these models, the success of a composition, at a given threshold, might tilt probabilities in the subgraphs of many parents, not just one as in the minimal architecture used here. The spreading exploration of subgraphs, mediated through parent parts via tilting, would in principle terminate by signaling every conjunctive hierarchy in which the root part has a conditional probability (or, more accurately, a localized approximation) greater than threshold. But to achieve this ideal, a number of modeling hurdles would need to be overcome, not the least of which is the development of accurate data models when multiple latent variables have overlapping supports. (See for example the POP model by Amit and Trouvé 2007, which includes a rigorous treatment of overlap in a generative framework.)

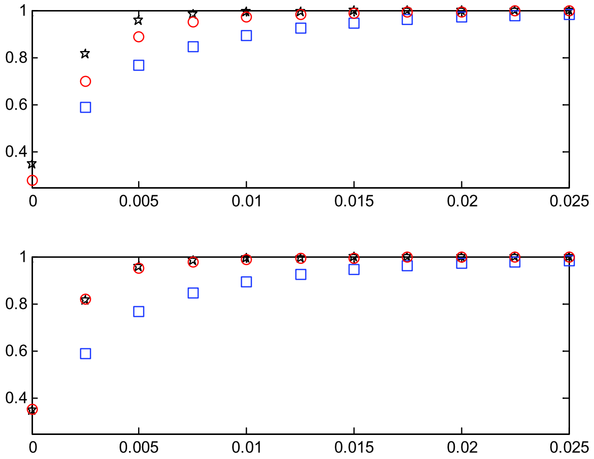


Fig. 9 Parts-Based Testing in a Hierarchical Model. The model depicted in Fig. 8 was used to generate conditional samples given $\mathcal{O} = 0$ and $\mathcal{O} = 1$, and an ROC curve was built for each of the four statistics, $\mathcal{S}_G$ (optimal, *black stars*), $\mathcal{S}_T$ (template, *blue squares*), $\mathcal{S}_P$ (parts, *red circles* in *top panel*), and $\mathcal{S}_Q$ (saliency, *red circles* in *bottom panel*). The relative merits of the four approaches are consistent with the earlier experiments (compare to Figs. 4 through 7), and with the analytic results discussed in Sects. 2 and 3. Saliency boosts performance, as compared to fixed-order parts-based testing, and parts-based testing generally outperforms template (all-or-nothing) testing. (Color scheme same as in Fig. 4)

We close this section with a remark about parallel versus serial computing. We have emphasized efficient sequential testing, but there are many opportunities for efficient parallel computation within the same parts-based testing paradigm. A simple example, for the execution of saliency-based testing, would be the assignment of a processor to each of the three subpart-to-part conjunctions in the architecture depicted in Fig. 8. Since the first step is to explore each of the parts Z_1, Z_2, and Z_3, and choose the most salient among those that exceed threshold (if any), and since these explorations involve no interaction among the parts, there is no reason not to explore them simultaneously. Following selection of the most salient part, the process repeats under the suitably tilted distribution on the remaining parts, offering the same opportunity for parallel computation. Not much would be gained in the minimal model used for our illustrations, but a great deal would be gained in more elaborate hierarchical models.

5 Summary and Conclusion

We have given mathematical and experimental evidence that in a high-resolution ("foveal") limit a sequence of local tests for the parts of an object can perform nearly optimal detection. If the sequence of tests is ordered by the local conditional probabilities of the parts ("saliency"), then performance can be essentially indistinguishable from optimal.

These results speak to a dilemma in compositional modeling: The choice between the presence and absence of a single object demands knowledge about the appearance of scenes both when the object is present *and* when it is absent. The absence of an object is easy to declare if nothing in the scene resembles the object. False detections in artificial vision systems occur overwhelming in regions that contain aspects or parts of an object, either in isolation or in compositions of other objects, but not the object itself. Therefore, to the extent that an object is made from the same components as other objects, i.e. to the extent that the world is compositional, a large class of objects needs to be modeled to recognize a single or small collection of objects.

We are suggesting, to the contrary, that in a compositional world a hierarchical model for an object provides an adequate model for the absence of the object, in terms of its parts and subparts, or more generally in terms of the subtrees of the hierarchy. We propose testing strategies that focus on tests for elementary parts and, recursively, on tests for subtrees.

Sequential testing of components is a form of coarse-to-fine search that can be extremely efficient when compared to global search. But context, generally acknowledged to be essential to good performance, is not local. We have demonstrated that an algorithm can be local in its examination of the pixel data, but still contextual in that it accounts for relationships among components. These relationships bias the interpretation of local data in accordance with the object being tested and the history of already successful tests. We have devised a statistic for sequential testing that is always between zero and one and is monotone decreasing with each new test, allowing for a single threshold and early stopping.

Broadly, we have discussed two diverse approaches to exploring the posterior distribution. Integration (i.e. summation, as in Sect. 3) focuses on identifying objects that are present with sufficiently high probability, regardless of instantiation. Alternatively, computation can be directed towards identifying specific instantiations of specific objects that are, by themselves, sufficiently likely, as illustrated in the experiment reported in Sect. 4. The best strategy might be a mixture of the two. To the extent that the supports of lowest-level parts are small and local, and progressively larger for higher-level parts, it might make sense to integrate low-level variables but look for specific instantiations of high-level variables. Thinking in terms of pose, the exact location of the most elementary parts may be of no interest and, in any case, highly ambiguous. Besides, the associated probabilities would depend critically on discretization, inviting unintended biases. A compelling argument can be made for integration. On the other hand, the states of high-level variables might code coarse features of location and structure, which could be of interest, or even vital, in particular applications. The right computation is very much architecture and application specific.

Appendix A

Proof of the Theorem We will write p_1 for the data model that generates the intensity at pixel $s \in S^1$ or $s \in S^2$ when $Z_1 = 1$ or $Z_2 = 1$, respectively, and p_0 for the data model at $s \in S \setminus (S^1 \cup S^2)$, or for $s \in S^1$ or $s \in S^2$ when $Z_1 = 0$ or $Z_2 = 0$, respectively. These models, p_0 and p_1, can be densities or probability mass functions. But they are assumed to be distinct, $p_0 \neq p_1$.

The theorem, as stated in Sect. 2, is for the special case $p_0(x_s) = G(x_s; 0, 1)$ and $p_1(x_s) = G(x_s; 1, 1)$, but we will assume only that $E_0 e^{tR} < \infty$ and $E_1 e^{tR} < \infty$, where $R = \log \frac{p_1(X)}{p_0(X)}$, and where E_0 is expectation with respect to $X \sim p_0$ and E_1 is with respect to p_1 (though less restrictive hypotheses could be used). The conditions hold, for example, in the Gaussian case as well as for many other densities, and also for probability mass functions p_0 and p_1 that have finite and common support, as well as in many discrete cases with infinite support.

For any statistic $\mathcal{S}(x_S)$

$$\mathcal{A}_{\mathcal{S}} = Prob\{\mathcal{S}(X_S) < \mathcal{S}(\tilde{X}_S) \mid X_S \sim P_1, \tilde{X}_S \sim P_0\}$$

where $P_1(x_S) = P(x_S \mid \{Z_1 = 1\} \cap \{Z_2 = 1\})$ and $P_0(x_S) = P(x_S \mid \{\{Z_1 = 1\} \cap \{Z_2 = 1\}\}^c)$. Let

$$\tilde{\epsilon}_0 = P(Z_1 = 0, Z_2 = 0 \mid \{\{Z_1 = 1\} \cap \{Z_2 = 1\}\}^c),$$

$$\tilde{\epsilon}_1 = P(Z_1 = 1, Z_2 = 0 \mid \{\{Z_1 = 1\} \cap \{Z_2 = 1\}\}^c)$$

and

$$\tilde{\epsilon}_2 = P(Z_1 = 0, Z_2 = 1 \mid \{\{Z_1 = 1\} \cap \{Z_2 = 1\}\}^c).$$

From (1):

$$\begin{aligned} P_0(x_S) &= \tilde{\epsilon}_0 P(x_S \mid Z_1 = 0, Z_2 = 0) \\ &\quad + \tilde{\epsilon}_1 P(x_S \mid Z_1 = 1, Z_2 = 0) \\ &\quad + \tilde{\epsilon}_2 P(x_S \mid Z_1 = 0, Z_2 = 1) \\ &= \tilde{\epsilon}_0 \prod_{s \in S^1} p_0(x_s) \prod_{s \in S^2} p_0(x_s) \prod_{s \in S \setminus (S^1 \cup S^2)} p_0(x_s) \end{aligned}$$

$$+\tilde{\epsilon}_1 \prod_{s\in S^1} p_1(x_s) \prod_{s\in S^2} p_0(x_s) \prod_{s\in S\setminus(S^1\cup S^2)} p_0(x_s)$$

$$+\tilde{\epsilon}_2 \prod_{s\in S^1} p_0(x_s) \prod_{s\in S^2} p_1(x_s) \prod_{s\in S\setminus(S^1\cup S^2)} p_0(x_s)$$

For brevity and clarity, we will focus on the special case $|S^1| = |S^2| = n$ (i.e. the parts have equal numbers of pixels in their supports). More generally $n^1 = |S^1|$ and $n^2 = |S^2|$, and in the foveal limit $\frac{n^1}{n^2} = \frac{n^1(n)}{n^2(n)} \to c$, $0 < c < \infty$, as $n \to \infty$. The proof is essentially the same.

Let $a_1^n = (a_1, a_2, \dots, a_n) = X_{S^1}$, $b_1^n = (b_1, b_2, \dots, b_n) = X_{S^2}$, $\tilde{a}_1^n = (\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_n) = \tilde{X}_{S^1}$, and $\tilde{b}_1^n = (\tilde{b}_1, \tilde{b}_2, \dots, \tilde{b}_n) = \tilde{X}_{S^2}$. Let $\epsilon_0 = P(Z_1 = 0, Z_2 = 0)$, $\epsilon_1 = P(Z_1 = 1, Z_2 = 0)$, and $\epsilon_2 = P(Z_1 = 0, Z_2 = 1)$. The proof of the theorem is based on the following lemma, whose proof we will return to shortly:

Lemma *Define the random variables*

$$v_i = \log\left(\frac{p_0(b_i)p_1(\tilde{b}_i)}{p_1(b_i)p_0(\tilde{b}_i)}\right), \qquad w_i = \log\left(\frac{p_0(a_i)p_1(\tilde{a}_i)}{p_1(a_i)p_0(\tilde{a}_i)}\right)$$

for $i = 1, \dots, n$, *where* $b_1^n \sim p_1$ *(iid),* $\tilde{b}_1^n \sim p_0$ *(iid),* $a_1^n \sim p_1$ *(iid) and* $\tilde{a}_1^n \sim p_1$ *(iid), and where* b_1^n, $\tilde{b}_1^n$, a_1^n, *and* $\tilde{a}_1^n$ *are all independent. Then*

$$A_{\mathcal{S}_G} \geq 0.5(\tilde{\epsilon}_1 + \tilde{\epsilon}_2) Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v}) \geq -\bar{v}\right), \tag{57}$$

$$A_{\mathcal{S}_P} \leq (2\tilde{\epsilon}_0 + 3\tilde{\epsilon}_1 + 3\tilde{\epsilon}_2) \times Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v}) \geq -\bar{v} + \frac{1}{n}\log(c)\right), \tag{58}$$

and

$$A_{\mathcal{S}_T} \geq \tilde{\epsilon}_1 Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v} + w_i) \geq -\bar{v}\right), \tag{59}$$

where $\bar{v} = E(v_i)$ *and* $c \leq 1$ *is a constant. (Notice that* $\bar{v} < 0$ *by Jensen's inequality,* $Ew_1 = 0$ *by symmetry, and* $Prob(v_1 > 0) > 0$, *which follows from* $p_0 \neq p_1$ *together with the fact that* $Ee^{tv_1} < \infty$, *implying that* p_0 *and* p_1 *have the same support.)*

The theorem follows from the lemma via large deviation bounds. We will follow the notation and development in Bahadur and Rao (1960). For estimating

$$Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v}) \geq -\bar{v}\right),$$

let $\varphi(t) = Ee^{t(v_1 - \bar{v})}$ and $\psi(t) = e^{-(-\bar{v})t}\varphi(t)$. Since $\varphi(t) < \infty$ for all t and $Prob(v_1 - \bar{v} > -\bar{v}) > 0$, according to Bahadur and Rao (1960) there exists a positive $\tau < \infty$ such that

$$\psi(\tau) = \inf_{t\in R} \psi(t) \equiv \rho.$$

By Theorem 1 of Bahadur and Rao (1960),

$$Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v}) \geq -\bar{v}\right) = \frac{\rho^n}{\sqrt{n}}O(1). \tag{60}$$

Similarly, for estimating

$$Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v}) \geq -\bar{v} + \frac{1}{n}\log(c)\right),$$

let $\psi_n(t) = e^{-(-\bar{v} + \frac{1}{n}\log(c))t}\varphi(t) = e^{-\frac{1}{n}\log(c)t}\psi(t)$. Then, we have

$$\begin{aligned} &Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v}) \geq -\bar{v} + \frac{1}{n}\log(c)\right) \\ &= \frac{(\inf_{t\in R}\psi_n(t))^n}{\sqrt{n}}O(1) \leq \frac{\psi_n(\tau)^n}{\sqrt{n}}O(1) \\ &= \frac{e^{-\log(c)\tau}\psi(\tau)^n}{\sqrt{n}}O(1) = \frac{\rho^n}{\sqrt{n}}O(1). \end{aligned} \tag{61}$$

Therefore, by (60) and (61) and the lemma, $\frac{A_{\mathcal{S}_P}}{A_{\mathcal{S}_G}}$ is bounded.

Next, we consider

$$Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v} + w_i) \geq -\bar{v}\right).$$

Let

$$\tilde{\varphi}(t) = Ee^{t(v_1 - \bar{v} + w_1)} = \varphi(t)Ee^{tw_1},$$

and let $\tilde{\psi}(t) = e^{-(-\bar{v})t}\tilde{\varphi}(t) = \psi(t)Ee^{tw_1}$. Since $\tilde{\varphi}(t) < \infty$ for all t and $Prob(v_1 - \bar{v} + w_1 > -\bar{v}) > 0$, there exists a positive $\tilde{\tau} < \infty$ such that

$$\psi(\tilde{\tau}) = \inf_{t\in R} \tilde{\psi}(t).$$

Thus, we have

$$\begin{aligned} &Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v} + w_i) \geq -\bar{v}\right) \\ &= \frac{\tilde{\psi}(\tilde{\tau})^n}{\sqrt{n}}O(1) = \frac{\psi(\tilde{\tau})^n(Ee^{\tilde{\tau}w_1})^n}{\sqrt{n}}O(1) \\ &\geq \frac{\rho^n(Ee^{\tilde{\tau}w_1})^n}{\sqrt{n}}O(1). \end{aligned} \tag{62}$$

Now, since $\tilde{\tau} > 0$ and $Ew_1 = 0$, $Ee^{\tilde{\tau} w_1} > 1$ by Jensen's inequality. Therefore, by comparing (62) and (61), and by the lemma, we obtain that $\frac{A_{\mathcal{S}_T}}{A_{\mathcal{S}_P}} \to \infty$ exponentially fast. It remains to prove the lemma:

$$\begin{aligned}
A_{\mathcal{S}_G} &= Prob((\mathcal{S}_G(X_S))^{-1} \geq (\mathcal{S}_G(\tilde{X}_S))^{-1} \mid X_S \sim P_1, \tilde{X}_S \sim P_0) \\
&= Prob\Bigg(\epsilon_0 \prod_{i=1}^{n} \frac{p_0(a_i)p_0(b_i)}{p_1(a_i)p_1(b_i)} + \epsilon_1 \prod_{i=1}^{n} \frac{p_0(b_i)}{p_1(b_i)} + \epsilon_2 \prod_{i=1}^{n} \frac{p_0(a_i)}{p_1(a_i)} \geq \epsilon_0 \prod_{i=1}^{n} \frac{p_0(\tilde{a}_i)p_0(\tilde{b}_i)}{p_1(\tilde{a}_i)p_1(\tilde{b}_i)} \\
&\quad + \epsilon_1 \prod_{i=1}^{n} \frac{p_0(\tilde{b}_i)}{p_1(\tilde{b}_i)} + \epsilon_2 \prod_{i=1}^{n} \frac{p_0(\tilde{a}_i)}{p_1(\tilde{a}_i)} \,\Bigg|\, (a_1^n, b_1^n) \sim P_1, (\tilde{a}_1^n, \tilde{b}_1^n) \sim P_0 \Bigg) \\
&\geq (\tilde{\epsilon}_1 + \tilde{\epsilon}_2) Prob\left(\prod_{i=1}^{n} \frac{p_0(b_i)}{p_1(b_i)} \geq \prod_{i=1}^{n} \frac{p_0(\tilde{b}_i)}{p_1(\tilde{b}_i)} \,\middle|\, b_1^n \sim p_1, \tilde{b}_1^n \sim p_0 \right) \\
&\quad \times Prob\left(\prod_{i=1}^{n} \frac{p_0(a_i)}{p_1(a_i)} \geq \prod_{i=1}^{n} \frac{p_0(\tilde{a}_i)}{p_1(\tilde{a}_i)} \,\middle|\, a_1^n \sim p_1, \tilde{a}_1^n \sim p_1 \right)
\end{aligned}$$

The last inequality is because the set

$$\left\{ \prod_{i=1}^{n} \frac{p_0(b_i)}{p_1(b_i)} \geq \prod_{i=1}^{n} \frac{p_0(\tilde{b}_i)}{p_1(\tilde{b}_i)}, \prod_{i=1}^{n} \frac{p_0(a_i)}{p_1(a_i)} \geq \prod_{i=1}^{n} \frac{p_0(\tilde{a}_i)}{p_1(\tilde{a}_i)} \right\}$$

is contained in the set

$$\begin{aligned}
&\Bigg\{ \epsilon_0 \prod_{i=1}^{n} \frac{p_0(a_i)p_0(b_i)}{p_1(a_i)p_1(b_i)} + \epsilon_1 \prod_{i=1}^{n} \frac{p_0(b_i)}{p_1(b_i)} + \epsilon_2 \prod_{i=1}^{n} \frac{p_0(a_i)}{p_1(a_i)} \geq \epsilon_0 \prod_{i=1}^{n} \frac{p_0(\tilde{a}_i)p_0(\tilde{b}_i)}{p_1(\tilde{a}_i)p_1(\tilde{b}_i)} \\
&\quad + \epsilon_1 \prod_{i=1}^{n} \frac{p_0(\tilde{b}_i)}{p_1(\tilde{b}_i)} + \epsilon_2 \prod_{i=1}^{n} \frac{p_0(\tilde{a}_i)}{p_1(\tilde{a}_i)} \Bigg\}.
\end{aligned}$$

Now, since

$$Prob\left(\prod_{i=1}^{n} \frac{p_0(a_i)}{p_1(a_i)} \geq \prod_{i=1}^{n} \frac{p_0(\tilde{a}_i)}{p_1(\tilde{a}_i)} \,\middle|\, a_1^n \sim p_1, \tilde{a}_1^n \sim p_1 \right) = 0.5,$$

we have

$$\begin{aligned}
A_{\mathcal{S}_G} &\geq 0.5(\tilde{\epsilon}_1 + \tilde{\epsilon}_2) Prob\left(\prod_{i=1}^{n} \frac{p_0(b_i)}{p_1(b_i)} \geq \prod_{i=1}^{n} \frac{p_0(\tilde{b}_i)}{p_1(\tilde{b}_i)} \,\middle|\, b_1^n \sim p_1, \tilde{b}_1^n \sim p_0 \right). \\
&= 0.5(\tilde{\epsilon}_1 + \tilde{\epsilon}_2) Prob\left(\frac{1}{n} \sum_{i=1}^{n} (v_i - \bar{v}) \geq -\bar{v} \right),
\end{aligned}$$

which is (57).

Next, let

$$G_1(a_1^n) = \frac{\epsilon_1}{\epsilon_1 + (1 - \epsilon_1) \prod_{i=1}^{n} \frac{p_0(a_i)}{p_1(a_i)}}$$

and

$$G_2(b_1^n) = \frac{\epsilon_{2|1}}{\epsilon_{2|1} + (1 - \epsilon_{2|1}) \prod_{i=1}^{n} \frac{p_0(b_i)}{p_1(b_i)}}.$$

where $\epsilon_{2|1} = P(Z_2 = 1 \mid Z_1 = 1)$. Then,

$$\begin{aligned}A_{\mathcal{S}_P} &= Prob(\mathcal{S}_P(X_S) \le \mathcal{S}_P(\tilde{X}_S) \mid X_S \sim P_1, \tilde{X}_S \sim P_0)\\ &= Prob(G_1(a_1^n)G_2(b_1^n) \le G_1(\tilde{a}_1^n)G_2(\tilde{b}_1^n) \mid (a_1^n, b_1^n) \sim P_1, (\tilde{a}_1^n, \tilde{b}_1^n) \sim P_0)\\ &= \tilde{\epsilon}_0 I_0 + \tilde{\epsilon}_1 I_1 + \tilde{\epsilon}_2 I_2\end{aligned}$$

where

$$I_0 = Prob(G_1(a_1^n)G_2(b_1^n) \le G_1(\tilde{a}_1^n)G_2(\tilde{b}_1^n) \mid a_1^n \sim p_1, b_1^n \sim p_1, \tilde{a}_1^n \sim p_0, \tilde{b}_1^n \sim p_0),$$

$$I_1 = Prob(G_1(a_1^n)G_2(b_1^n) \le G_1(\tilde{a}_1^n)G_2(\tilde{b}_1^n) \mid a_1^n \sim p_1, b_1^n \sim p_1, \tilde{a}_1^n \sim p_1, \tilde{b}_1^n \sim p_0)$$

and

$$I_2 = Prob(G_1(a_1^n)G_2(b_1^n) \le G_1(\tilde{a}_1^n)G_2(\tilde{b}_1^n) \mid a_1^n \sim p_1, b_1^n \sim p_1, \tilde{a}_1^n \sim p_0, \tilde{b}_1^n \sim p_1).$$

Now

$$\begin{aligned}I_0 &\le Prob(G_1(a_1^n) \le G_1(\tilde{a}_1^n) \mid a_1^n \sim p_1, \tilde{a}_1^n \sim p_0)\\ &\quad + Prob(G_2(b_1^n) \le G_2(\tilde{b}_1^n) \mid b_1^n \sim p_1, \tilde{b}_1^n \sim p_0)\\ &= 2Prob\left(\prod_{i=1}^n \frac{p_0(b_i)p_1(\tilde{b}_i)}{p_1(b_i)p_0(\tilde{b}_i)} \ge 1 \;\middle|\; b_1^n \sim p_1, \tilde{b}_1^n \sim p_0\right)\end{aligned}$$

Next, since $G_1(\tilde{a}_1^n) \le 1$,

$$\begin{aligned}I_1 &\le Prob(G_1(a_1^n)G_2(b_1^n) \le G_2(\tilde{b}_1^n) \mid a_1^n \sim p_1, b_1^n \sim p_1, \tilde{b}_1^n \sim p_0)\\ &= Prob\Bigg(\epsilon_1(1-\epsilon_{2|1})\prod_{i=1}^n \frac{p_0(\tilde{b}_i)}{p_1(\tilde{b}_i)} \le \epsilon_{2|1}(1-\epsilon_1)\prod_{i=1}^n \frac{p_0(a_i)}{p_1(a_i)} + \epsilon_1(1-\epsilon_{2|1})\prod_{i=1}^n \frac{p_0(b_i)}{p_1(b_i)}\\ &\quad + (1-\epsilon_1)(1-\epsilon_{2|1})\prod_{i=1}^n \frac{p_0(a_i)}{p_1(a_i)}\prod_{i=1}^n \frac{p_0(b_i)}{p_1(b_i)} \;\Bigg|\; a_1^n \sim p_1, b_1^n \sim p_1, \tilde{b}_1^n \sim p_0\Bigg)\\ &\le J_1 + J_2 + J_3\end{aligned}$$

where

$$J_1 = Prob\left(\epsilon_1(1-\epsilon_{2|1})\prod_{i=1}^n \frac{p_0(\tilde{b}_i)}{p_1(\tilde{b}_i)} \le 3\epsilon_{2|1}(1-\epsilon_1)\prod_{i=1}^n \frac{p_0(a_i)}{p_1(a_i)} \;\middle|\; a_1^n \sim p_1, \tilde{b}_1^n \sim p_0\right),$$

$$J_2 = Prob\left(\epsilon_1(1-\epsilon_{2|1})\prod_{i=1}^n \frac{p_0(\tilde{b}_i)}{p_1(\tilde{b}_i)} \le 3\epsilon_1(1-\epsilon_{2|1})\prod_{i=1}^n \frac{p_0(b_i)}{p_1(b_i)} \;\middle|\; b_1^n \sim p_1, \tilde{b}_1^n \sim p_0\right)$$

and

$$J_3 = Prob\left(\epsilon_1(1-\epsilon_{2|1})\prod_{i=1}^n \frac{p_0(\tilde{b}_i)}{p_1(\tilde{b}_i)} \le 3(1-\epsilon_1)(1-\epsilon_{2|1})\prod_{i=1}^n \frac{p_0(a_i)}{p_1(a_i)}\prod_{i=1}^n \frac{p_0(b_i)}{p_1(b_i)} \middle| a_1^n \sim p_1, b_1^n \sim p_1, \tilde{b}_1^n \sim p_0\right)$$

When n is large enough,

$$\max(J_1, J_2, J_3) \le Prob\left(\prod_{i=1}^n \frac{p_0(b_i)p_1(\tilde{b}_i)}{p_1(b_i)p_0(\tilde{b}_i)} \ge c_1 \;\middle|\; b_1^n \sim p_1, \tilde{b}_1^n \sim p_0\right),$$

where

$$c_1 = \min\left(\frac{1}{3}, \frac{\epsilon_1(1-\epsilon_{2|1})}{3\epsilon_{2|1}(1-\epsilon_1)}\right)$$

is a constant. Therefore,

$$I_1 \leq 3Prob\left(\prod_{i=1}^{n} \frac{p_0(b_i)p_1(\tilde{b}_i)}{p_1(b_i)p_0(\tilde{b}_i)} \geq c_1 \;\middle|\; b_1^n \sim p_1, \tilde{b}_1^n \sim p_0\right)$$

Similarly,

$$I_2 \leq 3Prob\left(\prod_{i=1}^{n} \frac{p_0(b_i)p_1(\tilde{b}_i)}{p_1(b_i)p_0(\tilde{b}_i)} \geq c_2 \;\middle|\; b_1^n \sim p_1, \tilde{b}_1^n \sim p_0\right)$$

where c_2 is a constant. Putting everything together,

$$\begin{aligned} A_{\mathcal{S}_P} &\leq (2\tilde{\epsilon}_0 + 3\tilde{\epsilon}_1 + 3\tilde{\epsilon}_2)Prob\left(\prod_{i=1}^{n} \frac{p_0(b_i)p_1(\tilde{b}_i)}{p_1(b_i)p_0(\tilde{b}_i)} \geq c \;\middle|\; b_1^n \sim p_1, \tilde{b}_1^n \sim p_0\right) \\ &= (2\tilde{\epsilon}_0 + 3\tilde{\epsilon}_1 + 3\tilde{\epsilon}_2)Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v}) \geq -\bar{v} + \frac{1}{n}\log(c)\right) \end{aligned}$$

where $c = \min(c_1, c_2, 1) \leq 1$, which is (58).
Finally,

$$\begin{aligned} A_{\mathcal{S}_T} &= Prob\left(\prod_{i=1}^{n} \frac{p_0(a_i)}{p_1(a_i)}\frac{p_0(b_i)}{p_1(b_i)} \geq \prod_{i=1}^{n} \frac{p_0(\tilde{a}_i)}{p_1(\tilde{a}_i)}\frac{p_0(\tilde{b}_i)}{p_1(\tilde{b}_i)} \;\middle|\; (a_1^n, b_1^n) \sim P_1, (\tilde{a}_1^n, \tilde{b}_1^n) \sim P_0\right) \\ &\geq \tilde{\epsilon}_1 Prob\left(\prod_{i=1}^{n} \frac{p_0(b_i)}{p_1(b_i)}\frac{p_1(\tilde{b}_i)}{p_0(\tilde{b}_i)}\frac{p_0(a_i)}{p_1(a_i)}\frac{p_1(\tilde{a}_i)}{p_0(\tilde{a}_i)} \geq 1 \;\middle|\; a_1^n \sim p_1, b_1^n \sim p_1, \tilde{a}_1^n \sim p_1, \tilde{b}_1^n \sim p_0\right). \\ &= \tilde{\epsilon}_1 Prob\left(\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v} + w_i) \geq -\bar{v}\right) \end{aligned}$$

Which is equation (59), completing the proof. □

Appendix B

Parts-based Testing, Ordered by Saliency in a Hierarchical System We use the notation introduced in Sect. 4. $\mathcal{S}_Q$ is like $\mathcal{S}_P$, but with random ordering. The ordering is defined by four permutations on $\{1, 2, 3\}$: k_i, $1 \leq i \leq 3$, which gives the order of appearance of the three parts, Z_1, Z_2, and Z_3, and $m_{i,j}$, $1 \leq i, j \leq 3$, which gives the order of appearance of the three subparts $Y_{i,1}$, $Y_{i,2}$, and $Y_{i,3}$. In sequential testing, the testing of each part involves a sequence of tests of the subparts. The second part tested, for example, is Z_{k_2}, for which $Y_{k_2,m_{k_2,3}}$ is the third subpart tested.

Once given the visitation schedules, k and m, the statistic $\mathcal{S}_Q$ is defined recursively:

$$\mathcal{S}_Q(x_S) = \mathcal{S}_Q^{\mathcal{O}}(x_{S^{\cdot,\cdot}}; P(\mathcal{O} = 1), k) \tag{63}$$

where

$$\begin{aligned} &\mathcal{S}_Q^{\mathcal{O}}(x_{S^{\cdot,\cdot}}; \alpha, k) \\ &\quad = \frac{\alpha}{\alpha + (1-\alpha)(.1)^3} \\ &\qquad \times \prod_{i=1}^{3} \mathcal{S}_Q^{Z_{k_i}}\left(x_{S^{k_i,\cdot}}; \frac{\alpha + (1-\alpha)(.1)^i}{\alpha + (1-\alpha)(.1)^{i-1}}, m\right) \end{aligned} \tag{64}$$

$$\begin{aligned} &\mathcal{S}_Q^{Z_i}(x_{S^{i,\cdot}}; \beta, m) \\ &\quad = \frac{\beta}{\beta + (1-\beta)(.2)^3} \\ &\qquad \times \prod_{j=1}^{3} \mathcal{S}_Q^{Y_{i,m_{i,j}}}\left(x_{S^{i,m_{i,j}}}; \frac{\beta + (1-\beta)(.2)^j}{\beta + (1-\beta)(.2)^{j-1}}\right) \end{aligned} \tag{65}$$

and finally

$$\mathcal{S}_Q^{Y_{i,j}}(x_{S^{i,j}}; \gamma) = \frac{\gamma l_{i,j}}{\gamma l_{i,j} + (1-\gamma)} \tag{66}$$

The algorithm for computing the (data-dependent) visitation schedules (i.e. k and m) is summarized in the following pseudo-code:

```
$\alpha = .01$
$SI = \{1, 2, 3\}$
while $SI \neq \emptyset$

    $l = 4 - |SI|$
    $\beta = \frac{\alpha+(1-\alpha)(.1)^l}{\alpha+(1-\alpha)(.1)^{l-1}}$
    for $i \in SI$

        $SJ = \{1, 2, 3\}$
        while $SJ \neq \emptyset$

            $r = 4 - |SJ|$
            $\gamma = \frac{\beta+(1-\beta)(.2)^r}{\beta+(1-\beta)(.2)^{r-1}}$
            $\tilde{m}_{i,r} = \arg\max_{j \in SJ} \mathcal{S}_Q^{Y_{i,j}}(x_{S^{i,j}}; \gamma)$
            $SJ = SJ \setminus \{\tilde{m}_{i,r}\}$

        end (while $SJ \neq \emptyset$)

    end (for $i \in SI$)
    $k_l = \arg\max_{i \in SI} \mathcal{S}_Q^{Z_i}(x_{S^{i,\cdot}}; \beta, \tilde{m})$
    $m_{k_l,\cdot} = \tilde{m}_{k_l,\cdot}$
    $SI = SI \setminus \{k_l\}$

end (while $SI \neq \emptyset$)
```

References

Ahuja, N., & Todorovic, S. (2008). Connected segmentation tree—a joint representation of region layout and hierarchy. In *CVPR'08*.

Amit, Y., & Geman, D. (1998). A computational model for visual selection. *Neural Computation*, *11*, 1691–1715.

Amit, Y., & Trouvé, A. (2007). Pop: Patchwork of parts models for object recognition. *International Journal of Computer Vision 75*(2).

Amit, Y., & Trouvé, A. (2010). *The more you look the more you see: efficient resource allocation for curve tracking in noise images* (Technical Report). University of Chicago, Statistics.

Bahadur, R. R., & Rao, R. R. (1960). On deviations of the sample mean. *Annals of Mathematical Statistics*, *31*, 1015–1027.

Barlow, H. (1994). What is the computational goal of the neocortex? In C. Koch, & J. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 1–22). Cambridge: MIT Press.

Bengio, Y., & LeCun, Y. (2007). Scaling learning algorithms towards ai. In L. Bottou, O. Chapelle, D. DeCoste, & J. Weston (Eds.), *Large-scale kernel machines*. Cambridge: MIT Press.

Blanchard, G., & Geman, D. (2005). Hierarchical testing designs for pattern recognition. *Annals of Statistics*, *33*, 1155–1202.

Borenstein, E., & Ullman, S. (2002). Class-specific, top-down segmentation. In *ECCV '02: proceedings of the 7th european conference on computer vision-part II* (pp. 109–124). Berlin: Springer.

Burl, M. C., & Perona, P. (1998). Using hierarchical shape models to spot keywords in cursive handwriting data. In *CVPR*.

Chang, L. B. (2010). *Conditional modeling and conditional inference*. PhD thesis, Brown University, Division of Applied Mathematics.

Chen, Y., Zhu, L., Lin, C., Yuille, A., & Zhang, H. (2007). Rapid inference on a novel and/or graph for object detection, segmentation and parsing. In *NIPS*.

Epshtein, B., & Ullman, S. (2005). Feature hierarchies for object classification. In *ICCV'05*.

Felzenszwalb, P. F., & McAllester, D. (2010). *Object detection grammars*. University of Chicago, Computer Science TR-2010-02.

Fidler, S., & Leonardis, A. (2007). Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR'07*.

Fleuret, F., & Geman, D. (2001). Coarse-to-fine face detection. *International Journal of Computer Vision*, *41*, 85–107.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, *28*, 3–71.

Harrison, M. T. (2005). *Discovering compositional structure*. PhD thesis, Brown University, Division of Applied Mathematics.

Jin, Y., & Geman, S. (2006). Context and hierarchy in a probabilistic image model. In *CVPR'06* (vol. 2, pp. 2145–2152). New York: IEEE Press.

Kokkinos, I., Maragos, P., & Yuille, A. (2006). Bottom-up & top-down object detection using primal sketch features and graphical models. In *CVPR'06*.

Ommer, B., & Buhmann, J. M. (2007). Learning the compositional nature of visual objects. In *CVPR'07*.

Moreels, P., & Perona, P. (2008). A probabilistic cascade of detectors for individual object recognition. In *ECCV*.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*, 411–426.

Shieber, S. (1992). *Constraint-based grammar formalisms*. Cambridge: MIT Press.

Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2005). Learning hierarchical models of scenes, objects, and parts. In *IEEE international conference on computer vision*.

Viola, P., & Jones, M. J. (2001). Robust real-time face detection. In *Proc. ICCV01* (vol. II, p. 747).

Warren, W. H. (2010). Direct perception. In E. Goldstein (Ed.), *Encyclopedia of perception*. Thousand Oaks: Sage.

Wu, T. F., & Zhu, S. C. (2010). A numerical study of the bottom-up and top-down inference processes in and-or graphs. *International Journal of Computer Vision*. doi:10.1007/s11263-010-0346-6.

Zhang, W. (2009). *Statistical inference and probabilistic modeling in compositional vision*. PhD thesis, Brown University, Division of Applied Mathematics.

Zhu, S. C., & Mumford, D. (2006). A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, *2*(4), 259–362.