Comparing Stories

Commentary on 'Experiencing and perceiving visual surfaces,' by K. Nakayama and S. Shimojo

Allan D. Jepson
University of Toronto

March 14, 1995

Nakayama and Shimojo consider[1] the visual interpretation of sparsely structured stereo images. The critical issue, common throughout this text, is how different interpretations for the same image should be compared. To do such comparisons the authors appeal to the notion of accidental views of a scene. The motivation for this is primarily scepticism that the visual system could gain, through experience, any accurate quantitative information about the prior probabilities of different scene interpretations. Indeed the generic view approach ignores the prior probabilities and treats every scene interpretation as equally likely a priori. I am sympathetic to this concern about the availability of accurate priors, especially considering the range of different and typically unknown contexts our visual systems must face, including various artificial contexts presented in psychophysics laboratories. But, as I discuss below, the generic view approach has some serious short comings, some of which can be traced back to this decision to ignore the priors.

Finally, I argue that there is a middle ground, which seems to me to be compatible with Nakayama's and Shimojo's general intuition, but does involve the use of some prior information. The critical point here is that the approach I suggest requires only rather general, qualitative properties of the various prior distributions. This latter part of the commentary is intended to be a carrot, enticing researchers such as the Nakayama and Shimojo to put a little structure in their priors (and just a little). But first we need the stick.

1 Against the generic view principle

The principle of generic views does not, by itself, explain the preferences for the interpretations of several stereo displays in the target article. There is a technical loophole in the

[1]This commentary is to appear in the book "Perception as Bayesian Inference" edited by D. Knill and W. Richards. The specific chapter it discusses is similar to Nakayama & Shimojo, Science, V.257, 1992, pp. 1357-1363.

argument which is due to the absence of a precise definition for what constitutes an accidental view or, conversely, a generic view. As we mention below this particular loophole could easily be plugged, but the more serious issue of the lack of a precise and well-founded definition remains.

1.1 What is a generic view?

It is useful to formalize the notion of a generic view. To do this the critical concept involves the set of images, say $E(I)$, which are considered to have the equivalent structure as some given image I. In particular, the set of all possible images is to be partitioned into equivalence classes, with any two images from the same equivalence class having the same structure. According to the target article, such a partitioning may be done on the basis of various topological properties of images. Let us ignore, for the moment, the details of the definition of these equivalence classes.

Equivalent viewpoints of a 3D scene can now be defined through the use of such a notion of image structure. To do this, consider a given image I_0 and a given 3D scene structure S. We assume that I_0 is a possible image of the scene S in that, when we view S from some viewpoint v_0, we obtain the image I_0. The set of equivalent viewpoints, namely $V(S, I_0)$, is then defined to be the set of all viewpoints, v, such that the image of S from viewpoint v is in the same equivalence class, namely $E(I_0)$, as the original image I_0.

The last component required for a formal definition is a way to measure the size of various subsets in the space of viewpoints. For example, for orthographic projection the set of viewpoints can be represented by a sphere, with the projection direction represented by the vector from the center to a point on the sphere. Given a subset of viewpoints on this sphere it is standard to use the area as the appropriate measure of the size of the subset. Perspective and/or stereo views require a larger dimensional viewpoint space and, similarly, the associated uniform volume measure could be used. Given these ingredients we can now define a generic view.

Generic View Definition: Given a particular image I_0 and a 3D scene structure S, suppose v_0 is a viewpoint such that the image of S from viewpoint v_0 is just I_0. This viewpoint of S is said to be generic if the set of equivalent views, namely $V(S, I_0)$, has positive measure on the set of all possible viewpoints.

For a concrete example consider the image given in Figure 13.1A along with the two interpretations displayed in Figure 13.1C and D. I refer to these interpretations as the folded cross and the floating bars, respectively. In order for the viewer to see the folded cross as depicted in the stereogram in Fig 13.1, the viewer needs to be positioned in the plane of the horizontal edges of the cross. Suppose that the notion of equivalent images mentioned above is defined so that straight lines are preserved within any one equivalence class. In particular, any viewpoint of the folded cross which is not in the plane of the horizontal edges would show the horizontal bar to have non-colinear edges in the image and, as a result, such a viewpoint is not considered to be equivalent. Therefore we find that the set of equivalent views is confined to the plane of the horizontal edges of the folded cross and must have measure zero. So the conclusion is that the the folded cross interpretation involves a nongeneric (i.e., accidental) view, as desired.

But what about the floating bars interpretation, can we conclude that the viewpoint for this is generic? The issue is that we do not have a specification for the set of equivalence classes $E(I)$ and, moreover, some choices for $E(I)$ lead to the conclusion that the viewpoint of the floating bars interpretation is also accidental. For example, suppose I consider the points defined by the centers of the horizontal and vertical bars in the floating bars interpretation. Such a point can be obtained, for example, by the intersection of straight lines connecting the opposite corners of the rectangle. Given that the our equivalence classes preserve straight lines, and without explicit directions to the contrary, this could be a valid construct from image data. If we allow such a construct, then we should note that these center points coincide in the fused stereo image. But this means that the viewer must be situated on the line passing through the centers of these two bars, which amounts to an accidental view. Almost any perturbation of the viewpoint would change this topological property of the center points projecting to the same point in the cyclopean image, and therefore change the image equivalence class. As a result, we would have to conclude that the viewpoint is accidental. But then both interpretations involve accidental views, in which case the principle of generic image sampling has nothing to say about the preference of one interpretation over the other.

This same situation holds for the disk illusion shown in Figure 13.6, with the perception being that the viewer is on the line passing through the center of the cross and the center of the transparent disk. Therefore the view of this interpretation could again be considered accidental, and the generic view principle fails to account for the preference of the disk interpretation.

The glitch mentioned above relies on the use of the centers of the various parts in the scene. One could work around this by refining the definition of an accidental view so that part midpoints are not considered in the various equivalence classes $E(I)$. But without a theoretical foundation as to why such a definition is appropriate, such a step seems arbitrary. The question remains, which image features should take part in deciding topological changes, which ones shouldn't, what topology should be used, and why? For that matter, it also seems arbitrary to restrict the definition of accidental views to *topological*[2] changes in the image structure. Why not include categorical changes which are not captured by standard image topologies? For example, could one usefully consider the categories of acute, right, and obtuse angles at various V-junctions in the image? The main point is that, in order to proceed, we need to have a clear definition of the equivalence class, $E(I)$, to be associated with any given image I. Moreover, one would hope that such a definition could avoid becoming a set of seemingly arbitrary criteria, and instead be well motivated from first principles.

1.2 A probabilistic perspective

In order to compare this with Bayesian approaches, it is useful to first relate the generic view principle to a probabilistic form. To do this, suppose we are given an image I and two possible

[2]Here I am referring to the use of topology in the third paragraph of the section titled "Ecological Optics and the Importance of Viewing Position" of the target chapter.

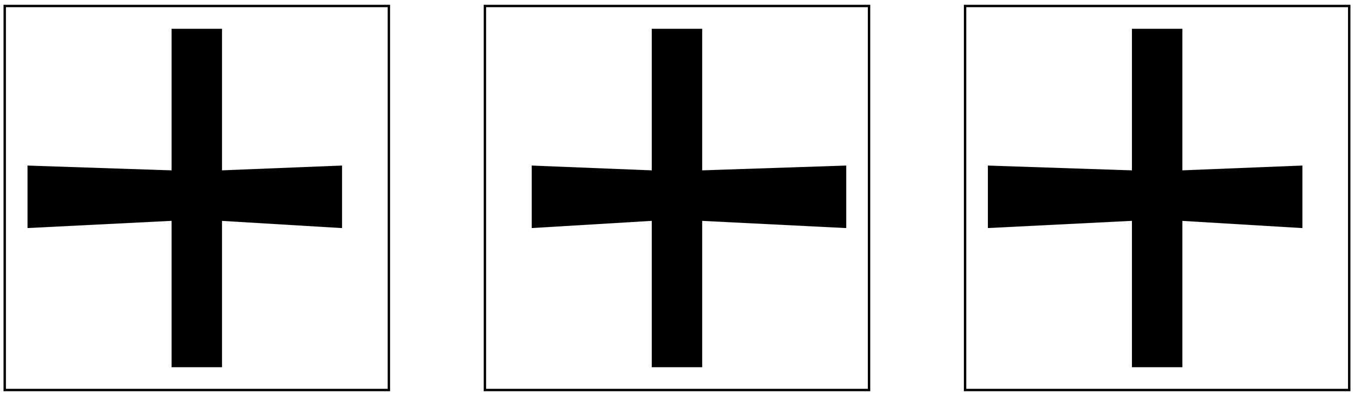

Figure 1.1: Divergers should use the right pair of stereo images, convergers the left. Observers report multiple, relatively discrete interpretations when this pair is viewed at arm's length with roving fixation.

scene structures S_1 and S_2. The principle of generic views considers only the likelihood of generating an image equivalent to I. That is, using the notation above for the set $E(I)$ of equivalent images, we consider the likelihoods $p(E(I)|S_i)$ for each scene interpretation S_i. The only random variable remaining in this expression is the viewpoint, which is taken to have a uniform prior. In fact, assuming that the viewpoint is selected independently of the scene S_i, this likelihood $p(E(I)|S_i)$ is just the prior probability of selecting a viewpoint v from the set of equivalent views $V(S_i, I)$. Moreover, given a uniform viewpoint distribution, this prior probability is proportional to the area of $V(S_i, I)$ on the viewing sphere. In these terms, the generic view principle dictates that we should prefer S_2 over S_1, say, whenever the likelihood $p(E(I)|S_2)$ given S_2 is positive while that given S_1 is zero. The principle can therefore be understood as special form of the maximum likelihood method applied to an extremely weak model of the world.

An alternative approach for choosing between two possible interpretations S_1 and S_2 is to compare their posterior probabilities given the observed image I. Note that the viewpoint is not included in the scene structures S_1 and S_2, rather it is being treated as a variable which we are not attempting to estimate accurately (see Freeman's notion of a generic variable in Chapter 12). By Bayes rule it follows that comparing the posterior probabilities is equivalent to comparing the products of the likelihood and the prior, as in $p(I|S_i)p(S_i)$. Note that there is no need here for the contentious set of equivalent images $E(I)$, but now we do require an estimate of the prior probability of the scene S_i. As we discussed at length in Chapter 4, the inclusion and qualitative form of this prior probability term can be critical in determining which interpretation is more probable given an observed image. We illustrate this in the following two subsections.

1.3 No accidental view, no preferences

One consequence of neglecting the prior probabilities is that the generic view principle does not account for many of the preferences exhibited by observers in the absence of accidental views. That is, observers can exhibit preferences for various interpretations without any accidents in the viewpoint position. For example, consider a modified version of the cross display, as depicted in Figure N.1. Here I have destroyed the colinearity (and the parallelism) in the edges of the horizontal bar. This colinearity was the critical factor, according to the argument in the target chapter, for determining that the viewpoint of the folded cross interpretation was accidental. For the display in Figure N.1 observers report up to three interpretations (although most observers do not report all cases, and many complain that the percepts are quite fleeting). One interpretation is similar to the folded cross, but with tapered wings. A second interpretation is similar to the floating bars interpretation with a flat horizontal bar, but now it is shaped like a bow tie in outline. The third interpretation is again similar to the folded cross, but with the vertical bar separate from and in front of the two receding wings.

The issue here is not so much why are these three interpretations distinguished, but rather why don't we get a continuum of possible interpretations? For example, we could parameterize a family of interpretations by the depth of the middle point of the horizontal bar. For all these solutions, the argument in the target chapter would conclude that the viewpoint is generic. Thus the generic viewpoint principle would not distinguish any of these possibilities. However, a central theme of this text is that such preferences can be represented within a Bayesian framework according to various prior models the observer has about the world. For example, the above three interpretations could arise from priors which distinguish flat surfaces, connected surfaces, and rectangular surfaces (respectively), as structures which occur relatively frequently in our world. The point here is simply that the generic view principle is clearly not enough on it's own to explain how we see; some other principles and preferences would need to be added. Presumably such preferences are also active in cases for which there is an accidental view, and a more complete theory would need to spell out how the various preferences interact with the generic view principle. In contrast, the Bayesian story emphasized elsewhere in this text, and discussed further below, attempts to gain a more unified perspective.

1.4 A preferred but accidental view

A related difficulty, caused again by ignoring the priors, is that some implausible interpretations end up being preferred according to the generic view principle. For example, the standard perception of the blocks world figure given in my commentary on Chapter 11 involves a highly accidental view. But there is an alternative interpretation of the scene for which all the colinear line segments may be taken as colinear in the world. In this interpretation various blocks must be floating in space, in special alignments, and at least one of the two cylinders must be distorted. Nevertheless, the view for this "floating blocks" interpretation is generic, while the common perception involves an accidental view. According to the principle of generic views, then, one should prefer the "floating blocks" interpretation over the standard perception of this figure. However, our visual systems do the opposite, and thereby violate the generic view principle in this example.

It is worthwhile to consider why the generic view principle fails in this example. The difficulty rests with ignoring the prior probability for the "floating blocks" interpretation. Given a simple qualitative model with smooth prior distributions for the viewpoint and for blocks floating freely in space, one can conclude that the posterior odds are strongly against any interpretation which involves the blocks aligned. Rather, the odds favour the interpretation that the viewpoint is accidental. The reason is that, even though the likelihoods favour the non-accidental view, the prior probabilities are more extremely skewed in favour of the other interpretation. As a result, the posterior probabilities turn out to be strongly against the non-accidental view. (The necessary calculations are similar to the discussion in Chapter 4, and are omitted here.) This is in accord with the standard perception of the blocks world figure.

The addition of a prior probability distribution may cause some concern. For example, the authors of the target article argue that it is implausible that the visual system has access to detailed knowledge of the prior probability distribution. I agree with this. But, as we discuss in Chapter 4, qualitative properties of the prior distribution can be sufficient to reliably select one interpretation over another. Moreover, since this required qualitative structure is rather simple, it seems plausible that our visual systems could learn it and make use of it. To illustrate this point further I outline in the next section a qualitative Bayesian formulation for the situation depicted in Figure 13.1.

2 Modal analysis

My emphasis in the discussion below is on the mathematical properties of inferences that can be made given a qualitative model of the prior probabilities for situations such as those depicted in the stereo examples of the the target chapter. The goals are to sketch the application of modal analysis to such a domain and to study the resulting inference problem. In other words, if we postulate that our perceptual systems perform Bayesian inference, and have priors of the form described below, then we can mathematically derive a certain set of results. Our goals here are to illustrate the approach, spell out the results, but not necessarily validate these postulates about our visual systems.

Consider a qualitative probabilistic model of the scene and the viewpoint geometry for situations which encompass the stereo examples in the target chapter. In particular, we consider scene models of the form described in Chapter 4, with the various prior distributions specified by qualitative probability distributions. Following my commentary on Chapter 11, we might frame such a prior in terms of a modal workshop. The intuitive idea is that the workshop can generate rectangular strips, fold them, hang them in place, and so on.

The important observation is that it is natural to assume that the prior distributions for various operations are modal. For example, a polygonal part may be constructed over a wide range of shapes, which might be modeled by a smooth distribution on the set of vertices. However, in addition to this smooth distribution, polygonal parts may be constructed in a variety of special ways, such as rectangular, symmetrical, or square. These latter shapes are represented by lower dimensional subsets of the overall imbedding space of polygons and, since they have nontrivial prior probability of occurring, we see that an appropriate prior distribution has the original smooth distribution in addition to various delta function distributions spread over these lower dimensional sets. Other examples of modal distributions occur with the relative positioning of two parts in space (where possible modes of positioning include various alignments of the part axes), and with the place and manner parts are folded (with modes in which straight folds made parallel or in line with other structures). In addition to qualitative priors for the various operations of the workshop, we also require a qualitative prior distribution for the viewer position. Here a natural prior is to assume the viewpoint is independent of the object in the scene and to use a smooth distribution over the space of viewpoints.

In Chapter 4 we define a context C to be a set of such prior models, which includes a specification of the various modes that are expected to occur, along with various non-degeneracy conditions. We avoid the details of such a specification here and instead concentrate on the various inferences that are sanctioned once a context has been chosen.

2.1 Defining accidents

Given a context such as the one described above, we can define a notion of one scene being accidental relative to another. This is quite different from the generic view approach described above, in that it is addressing the relative sizes of the priors for two scene interpretation, say $p(S_1|C)$ and $p(S_2|C)$, without regards to an image or an observer. An accidental scene is easiest to illustrate by suppressing one of the modes we have already assumed to exist. For example suppose we modified the context such that, when objects were floating in space their relative positions did not have any alignment modes, but rather consisted of only a smooth distribution. Let us denote this modified context by C_m instead of C. For context C_m the orthogonal alignment of the floating bars would have to arise by chance, while that of the folded cross could be explained by the remaining modes of the context (i.e. cuts can be made orthogonally to other cuts, and folds can be made along the edge of the vertical bar). Thus, in such a context, the prior probability of the floating bars interpretation, say $p(S_2|C_m)$, is vanishingly small relative to that of the folded cross (here, as in Chapter 4, we are considering the limit as an error tolerance goes to zero.) It is convenient to express this in terms of the ratio of priors, as $p(S_2|C_m)/p(S_1|C_m) \to 0$. As a result, this floating bars interpretation is said to be accidental relative to the folded cross in this modified context.

For the original context, C, there are sufficient modes to account for the various structures in both the folded cross, S_1, and the floating bars, S_2. As a result, the priors $p(S_1|C)$ and $p(S_2|C)$ are comparable. That is, the prior ratio does not go to zero, or grow unbounded, as the error tolerance is refined. Therefore, in the context C, neither interpretations is (a priori) accidental relative to the other.

Another example of an accidental structure occurs in this original context C when a new piece of data is considered. Suppose that we are told that the distance from one tip to the other in the horizontal bar, in either interpretation, has to be just the same as the length of the vertical bar. To account for this observation using a folded cross interpretation, the length of the folded horizontal segments and the fold angle need to be chosen in such a way as to arrive at the correct length. The prior probability for doing this, in our current context, turns out to be vanishingly small (again, in the limit as the error tolerance goes to zero). However the same property can be explained in the floating bars interpretation by appealing to the mode that the two bars are cut into the same shape. As a result the folded cross has a vanishingly small posterior probability relative to the floating bars, given this additional observation, and would therefore be considered to be accidental relative to the floating bars.

The effect of being given an image instead of a single observation can be viewed in a similar way. That is, given the two scene interpretations S_1 and S_2, we wish to compare their posterior probabilities given the context C and the image I, namely $p(S_1|I,C)$ and $p(S_2|I,C)$. We say S_1 is accidental relative to S_2, given I and C, if $p(S_1|I,C)/p(S_2|I,C)$ is vanishingly small in the limit as the error tolerance goes to zero. In such a case, it is appropriate to prefer the story told by S_2 over that of S_1, at least for sufficiently small errors in this context C. Recall from Chapter 4 that the ratio of posterior probabilities is just the product of the ratio of likelihoods with the ratio of the prior probabilities. That is

$$\frac{p(S_1|I,C)}{p(S_2|I,C)} = \frac{p(I|S_1,C)}{p(I|S_2,C)} \frac{p(S_1|C)}{p(S_2|C)},$$

where the first term on the right hand side is the likelihood ratio and the second term is the ratio of prior probabilities.

How does this formulation work out for the stereo display in Figure 13.1? The answer turns out to be interesting in that it depends on how well the viewpoint can be estimated from the observed position of the vertical bar and the ends of the horizontal bar. Note that these positions are common to the folded cross and floating bars, and thus any estimate obtained from just this data is not dependent on the choice between the two. There are two possible scenarios.

In the first case we assume that the information available from the disparity of the vertical bar, and the endpoints of the horizontal bar is sufficiently accurate to narrowly constrain the viewpoint around the fronto-parallel direction. Moreover, the uncertainty in the viewpoint position is sufficiently small that we cannot reliably resolve the distinction between the folded cross and the floating bars from a significant fraction of viewpoints within this range. From this property alone it follows that the likelihoods, $p(I|S_1,C)$ and $p(I|S_2,C)$, are comparable. Since the posterior probability ratio is equal to the product of the likelihood ratio times the prior probability ratio, it follows that the posterior probability ratio is essentially the same as the prior ratio. As discussed above, this ratio is not extreme in the context C. Therefore, in this case, no information is gained about which of the two interpretations, S_1 or S_2, should be preferred!

In the second case, we assume that we cannot reliably estimate the viewpoint position. Moreover, the derived set of possible views is sufficiently broad so that only a small fraction are consistent with the folded cross. This is the classic case in which the generic view assumption applies. Here the likelihood ratio is determined by the fraction of views which are consistent with the folded cross interpretation, and we are assuming that this ratio strongly favours the floating bars interpretation. Also, because the prior ratio is not extreme, we can in this case reliably prefer the floating bars interpretation.

One difference between these calculations and the standard generic view assumption is that image data is being taken into account to first limit the set of viewpoints. (The computations are similar to those involved in the 'generic view' terms described in Chapter 12.) The appropriate likelihood ratio depends on this limited set of viewpoints, not on the overall space of views. If this limited set of views is sufficiently small we may gain no information regarding the choice of S_1 or S_2. Various quantities can be expected to have an effect on the accuracy of the estimate of the viewpoint, such as the visual angle of the display and the presence of other structures such as the bounding boxes in Figure N.1. These are testable predictions of the qualitative probabilistic model.

3 Telling stories

In a sense, the selected scene interpretation S_i and the context C together tell a story about how an image arose. This story goes considerably further than just the description of scene structure, provided by S_i, in that the context C includes a specification of various modes that are present in the current domain. The context C can be thought of as describing the 'modal workshop' (see my commentary on Chapter 11) which generated the scene; it specifies which world properties can be generated non-accidentally in the specific domain. In particular, in order to arrive at a percept, some simple qualitative properties of the prior distribution must be represented and used to compare different scene interpretations.

This picture of the perceptual system as a story teller raises several issues. The first of which is how the various modes are indexed in order to construct the particular context. For example, in the context C discussed above, the orthogonal alignment of various parts floating in space was assumed to be modal. Somehow such a mode must be recovered from a knowledge base of possible modes. Perhaps an associative memory could form the basis of a suitable indexing scheme for individual modes. A second issue is how the various modes in our environment may be learned, that is, how our knowledge base of modes is built up in the first place. This involves the construction of mixture models (see Chapter 4) for the probability density functions of various scene properties, and perhaps the work on soft competitive learning in the connectionist literature could be applied (for example, see [1]).

Finally, a third issue involves the apparent preference of some contexts over others. For example, many possible contexts could be assumed for Figure N.1 above. In particular, why not choose any subset of the modes in context C rather than the whole set? Suppose, for the sake of the argument, we eliminated folding from the context. This alone would rule out the folded cross interpretation. Alternatively, a different (but still seemingly arbitrary) assumption is that the context could lack the operation of aligning separately floating parts (see the discussion of C_m above). Such variations in context eliminate modes required to plausibly construct particular artifacts, and can thereby change which interpretations are most probable. Given this variability it is critical for a perceptual theory to provide a basis on which one context can be preferred over others.

The straight forward Bayesian approach would be to supply priors on these contexts, say $p(C_k)$. The estimation problem is then replaced by one in which one seeks the maximum a posteriori probability of both the scene interpretation and the context, namely $p(S_i, C_k|I)$.

By Bayes rule, we would only need to compare products of the form $p(I|S_i, C_k)p(S_i|C_k)p(C_k)$. Here the first two terms are the familiar image likelihood and the scene prior given the context C_k, while the latter term is the prior for the context C_k. This is appealing in terms of it's simplicity, but I find it daunting to begin to specify priors on all possible contexts. Certainly some sort of constructive process would be needed to compute a prior for a novel context, such as the various contexts described above. But what is the nature of such a constructive process? What sort of output can be expected; presumably not quantitative prior probabilities, but rather qualitative probabilities or alternatively just a preference ordering? What are the implicit assumptions and constraints behind such a constructive process? These are some of the issues raised in Whitman Richards' commentary on Chapter 7, although he does not take a Bayesian perspective.

The importance of context selection, and the difficulty in providing reasonable priors, is highlighted further when one considers enlarging the context to include things like communication conventions or the purpose of the people showing you the artifact. In terms of communication conventions, one argument might be that if the demonstrators wanted to communicate the folded cross then they would display it using a representative view. Roughly speaking, a representative view could be defined as a viewpoint from which our perceptual systems will typically recover the folded cross interpretation. The fronto-parallel stereo view is not representative for the folded cross, but it is for the floating bars. In fact, it might be ideal for the floating bars, in terms of communicating the shape of the two bars. Thus, if we assume a context in which the demonstrators are attempting to clearly communicate an artifact, then it make sense to select the floating bars interpretation. Of course an alternative purpose, and an alternative context, is that the demonstrators wish to show you an ambiguous view, maybe to see how you cope with it. But, you should ask, what is the prior probability of a context as convoluted as that?

References

[1] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1), 1991.