A Probabilistic Model of the Visual System

Siddhartha Kasivajhula
PSYCH221/EE362 Final Project

Abstract

A probabilistic model of the visual system is proposed. The model exhibits the following 3-phase high-level behavior: First, a prior distribution is generated over predicted features in the scene. Then, the eyes supply visual information (evidence). And finally, the distribution is modified based on the evidence received, and this posterior distribution again serves as the prior for the subsequent time-step. The process of observation is modeled by a Hidden Markov Model; and a 2-level *noisy-or* Bayes Net classifier is used to model the process of scene interpretation. Finally, the results of the Bruner and Potter Experiment are reinterpreted using this probabilistic model, and an explanation of the results is suggested.

1 INTRODUCTION

Conventional models of the visual system maintain that:

(**i**) The purpose of vision is to provide a fully detailed representation of the visual scene.

(**ii**) Higher levels in the hierarchical processing of visual information depend on lower levels, but not vice-versa [Churchland].

Churchland and Ramachandran, in "A Critique of Pure Vision," argue for a revision of this conventional model, and propose the following paradigm as a replacement for the current one:

(**i**) The purpose of vision is to represent the visual scene in a manner so as to optimize "the four Fs: feeding, fleeing, fighting, and reproduction." Complete scene representation is unnecessary for this purpose, and at any point in time, the visual scene is represented only partially – based on what is immediately relevant to the organism.

(**ii**) There is no strict hierarchy in visual information processing.

However, they do not suggest a possible formalism for representation of these ideas (and even note this fact). I propose such a framework, based on a probabilistic model of the visual system.

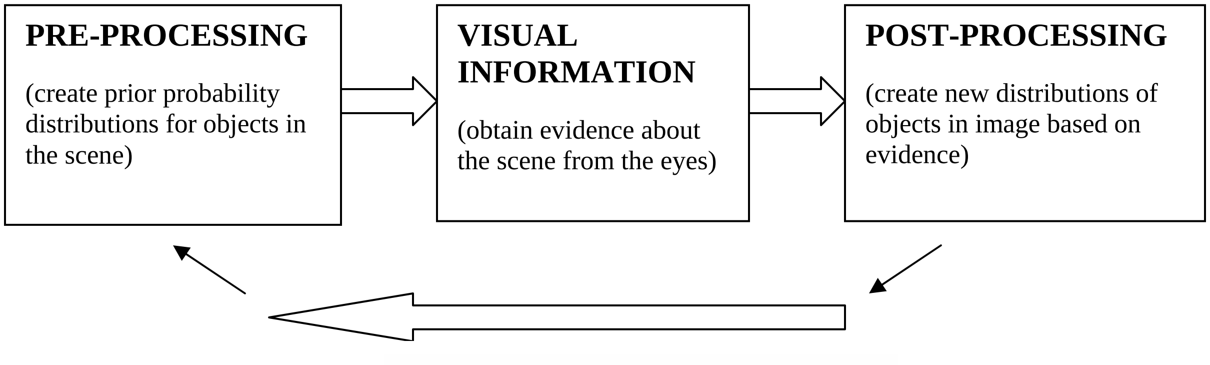
PRE-PROCESSING
(create prior probability distributions for objects in the scene)
VISUAL INFORMATION
(obtain evidence about the scene from the eyes)
POST-PROCESSING
(create new distributions of objects in image based on evidence)

Figure 1: Visual System Architecture

In particular, I suggest that the feedback mechanism shown in Figure 1 is always active in the visual system, and that this is the mechanism of "seeing."

2 THE MODEL

I. Representation

First, to build our framework for the model, we must decide upon how the visual scene is represented. I motivate a possible representation with the following:

1. Wang and Ross show that people are able to recognize familiar object much more quickly than unfamiliar objects [Wang]. *This implies that prior knowledge of objects is used in visual perception.*

2. Consider the following thought experiment: Imagine a white room filled with dull white objects. It would be impossible to have any sense of perspective in such a room, or to identify the objects in the room or where they're located. This thought experiment serves to illustrate that we are able to make sense of visual information only when there are *contrasts*. This makes sense for the following mathematical reason: whenever there is a contrast between two regions, the interface between the regions is perceptible as a geometric shape. Any discernable geometric shape can provide 3-D information about the scene, since relative motion between the shape and the observer results in the perceived deformation of the geometric shape. Lines may grow longer or shorter depending on the observer's perspective or proximity to the lines. But flat, featureless surfaces give no information apart from their color.

So we may conclude from this argument that there are certain elements of the visual scene that yield useful and comprehensible information, and that it should be possible to define such elements. We will refer to all such elements simply as "features," without attempting to rigorously define exactly what these may be. Intuitively, some possible features may be color, geometry.

If we assume that the size of the complete space of features (total number of possible features) is 'n', with features indexed by 'i', then we represent a **feature** as a vector of length n:

$$f = [0 \;\; 0 \dots 0 \;\; 1 \;\; 0 \dots 0 \;\; 0]$$

The vector will contain a single '1', with all the other entries set to 0. A '1' at index 'j' represents that feature 'j' is present. Let the set of all such recognizable features be called 'F'. Formally:

$$F = \{f_i \in R^n : 1 \leq i \leq n,\ n \in Z\}$$

An **observation** is defined to be some subset of the set of all features F. For convenience, we will define the set of all possible observations O as the *power set*[1] of F:

$$O = P(F),$$

which means an observation o is now defined to be a member of this set:

$$o \in O.$$

[1] The power set is the "set of all subsets".

An **object** is represented as a vector of *weights* that are associated with corresponding features:

$$\Psi: \Psi \in R^n, \Psi(i) = w_i; w_i \in (0, 1)$$

For example, if we assume a feature space of size 5, the following is a valid object vector:

$$\Psi = [0 \;\; 0.45 \;\; 0.9 \;\; 0.1 \;\; 0.6]$$

II. Observation

The process of observation can be described as sensing information from the visual scene. As described above, we model the scene as a set of feature vectors **S**. The act of observation can now be described as obtaining a noisy set o of feature vectors from S. Note that, since noise is introduced here, the set o may contain feature vectors *not* present in S. This corresponds to the observer seeing things that are not really there, that are just artifacts of atmospheric interference or unusual lighting. The process of observation as described here can be modeled as a Hidden Markov Model (HMM), as follows:

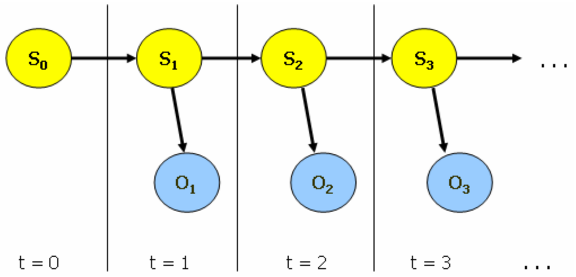

Figure 2: Hidden Markov Model

The model encodes the following behavior: the process of observation is broken up into discrete time steps. The "true" set of features in the scene at any time step is dependent on the features present in the scene at the previous time step (features in the scene at t are likely to be present in the scene at t+1, in some form). At each time step, an observation of the scene is made.

III. Interpretation

Recall that we had mentioned previously that prior knowledge of objects is used in scene interpretation. To capture this, I chose to model interpretation of the visual scene as a 2-layer *noisy-or* Bayesian Network, as shown below. The "noisy or" assumption is that the probability that some set of objects causes a feature to be observed is simply the probability that at least one of the objects caused it. This is a sensible assumption to make in this context since it is unlikely that a single feature in the scene is part of several different objects.

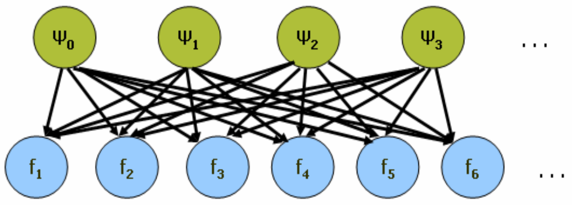

Figure 3: Noisy-or Bayes Net

The noisy-or model is used frequently in medical diagnosis, with "symptoms" replacing "features," and "diseases" replacing "objects." [Szolovits].

IV. Mechanism

The complete working of the model is depicted in Figure 4. A prior distribution is first created over features, and consequently objects, in the scene. An observation provides a set of features to the classifier, and the weights for the objects are recomputed conditioned on this new evidence. The weights over the entire set of objects are normalized to sum to 1 to make it a legitimate probability distribution. At any point in

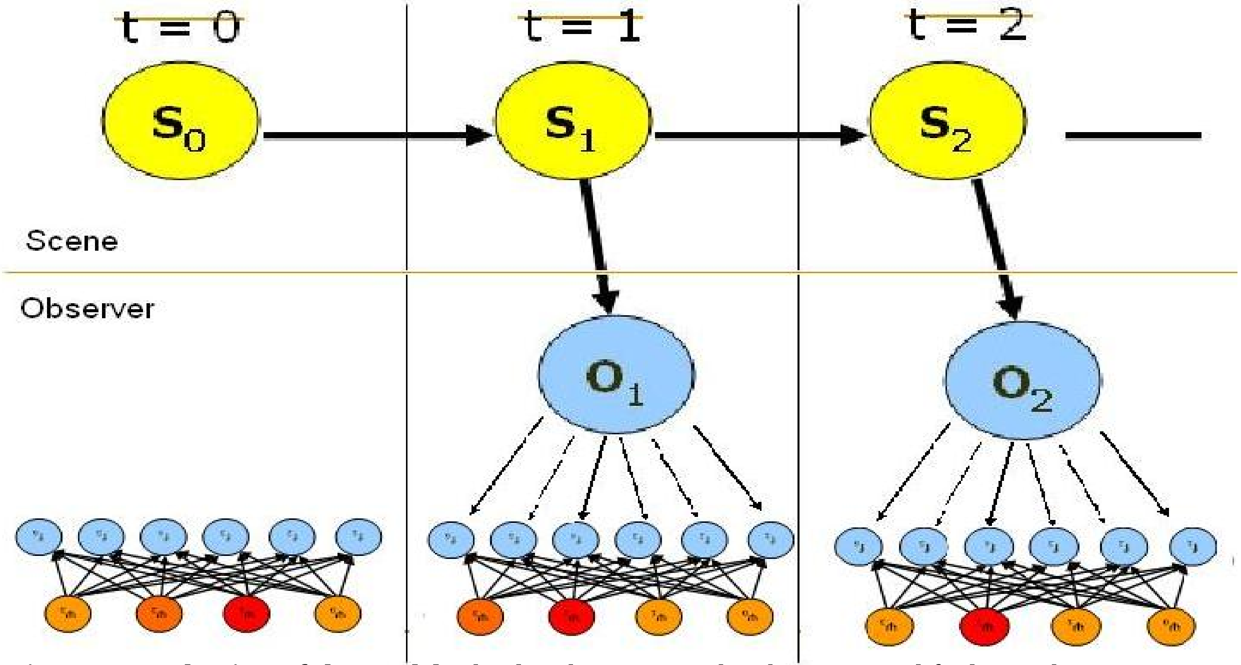

Figure 4: Mechanism of the Model. The distribution over the objects is modified at each time step when visual evidence is received. For illustration, darker shades of red represent higher weight for an object. (Note that the noisy-or Bayes net is upside down).

time, the object perceived by the observer is the object which has the highest weight in the distribution, Ψ_{max}. This process repeats for every timestep.

3 APPLICATION: THE BRUNER AND POTTER EXPERIMENT

Bruner and Potter in 1964 performed the following experiment: An image was brought from a state of complete defocus to complete focus over a period of 60 seconds. One test group was allowed to watch the image from t=0, a second group was brought in at t=30s, and a third was brought in at 45s. The surprising results of the experiment were that the people who came in later were able to identify the contents of the image *earlier* than those who had been looking at the image longer[Bruner].

We will now interpret these results using our model and provide an explanation.

If we assume, for simplicity, that a single object Ψ_s is contained in image, then the result of the experiment can be captured in the following probabilistic statement:

$$P(\Psi_{max} = \Psi_s \,|\, O_1, \ldots, O_t) < P(\Psi_{max} = \Psi_s \,|\, O_t)$$

That is, the probability of the observer's best-hypothesis object being the correct object is lower given all observations upto that point than just given the observation *at* that timestep.

The Nature of the Data

To explain this, let us consider the nature of the scene features S presented to the observer. At each subsequent timestep, the *same image* is shown to the observer, but more in-focus than in the previous timestep. This means that:

1. The information content of S_t is strictly greater than the information content of S_{t-1}.

2. The information content of S_t is a *superset* of the information contained in S_{t-1}

3. Any other difference between S_{t-1} and S_t is precisely equal to the additional noise present in S_{t-1}.

What this means in our model

Since observations O_t are direct children of the scene feature S_t in the HMM, the above three properties can be said to be true of O_t and O_{t-1}, as well. Following the flow in the model, we find that the features presented to the noisy-or scene classifier cause certain weights to be assigned to all of the hypothesis objects Ψ_i. Since the weights are modified iteratively at every timestep, there will tend to be a certain persistence of hypotheses – if, in the present observation, there is very little to support the presence of object Ψ_5 in the scene, the weight of Ψ_5 will be decremented. So if Ψ_5 had developed high weight until time t, it might still be relatively highly weighted in timestep t+1, despite lack of evidence to support its existence at time t.

Now, as per our discussion of the nature of the Bruner and Potter image across timesteps, we can use the 3 points noted above to conclude that the hypothesis weights assigned after looking at the image for t timesteps are fundamentally less accurate than the weights that would be assigned given only the image at time t, which is precisely the result that we set out to explain.

4 CONCLUSIONS AND FUTURE DIRECTION

This model is still quite simplistic when compared with the visual system. It serves to explain some experimental results, but there are many parameters that are still unmodeled. For example, when we see, we associate observed objects with a region (area) in our field of vision. This model does not consider feature locality at all. Additionally, there might be some other levels in the hierarchy in visual interpretation which are not captured here. For example, we are able to track objects moving through our field of vision even if we can't discern what they are – we are able to recognize the objects as entities distinct from their surroundings even before any higher level understanding of the objects is arrived at. This seems to suggest that *object delineation* is done *independently of* and *prior to* any object classification. If this is the case (that this level is independent of the classification level), then we could model this level and have it fit neatly on top of the model described here without having to change anything with this model.

One other probabilistic approach that could be explored as a possible model for aspects of the visual system is *particle filtering* – where several hypotheses are kept track of simultaneously and are all updated with observations at each time step. More likely hypotheses are duplicated and perpetuated across timesteps and less likely ones "die off." This method could be applied to depth perception, I feel. For a given object, we might maintain several hypotheses regarding the distance to this object: say the hypotheses are "5m away, 5.2m away, 4.9m away." Now if we move 0.3m closer to the object, all of these hypotheses will be modified to reflect this. So our new hypotheses will be "4.7m away, 4.9m away, 4.6m away." And at each step, less likely hypotheses will be replaced by more likely ones. So, for example, "4.6m away" might be unlikely, and could be replaced with "4.7m away". Our new hypotheses would then be "4.7m away, 4.9m away, 4.7m away." Intuitively, I feel this is a good model for depth perception, and it might be worth exploring.

5 ACKNOWLEDGEMENTS

I would like to thank Rajat Raina, CS PhD candidate at Stanford University, for advice and discussion; Prof. Daniel C. Richardson of UCSC for the video used in my presentation (video made by Itti, Koch at Caltech); and Manu Parmar and Prof. Wandell for advice and encouragement at the critical initial stages of the project.

6 REFERENCES

1. P.Churchland, et. al, "A Critique of Pure Vision"

2. L. Wang, J. Ross, "Interactions of neural networks: Models for distraction and concentration", 1990.

3. P. Szolovits. "Uncertainty and Decisions in Medical Informatics". Meth. Inform. Med. Vol. 34, No 1/2. 1995. Obtained online at: http://dsg.harvard.edu/courses/hst951/PDF/Uncertainity%20and%20Decisions%20in%20Medical%20Informatics.pdf

4. J. Bruner, M. Potter, "Interference in Visual Recognition", 1964.

5. A Doucet, N de Freitas and N Gordon. Sequential Monte Carlo Methods in Practice. Springer, 2001.

A.I APPENDIX I: MATLAB CODE

A MATLAB implementation of some aspects of this model is provided. The code is only meant for illustrative purposes, to demonstrate how this model might be implemented, and does not necessarily represent the best way to do so. Please read the README file contained in the submission for more information.

A.II APPENDIX II: AN EXPERIMENT

I conducted this experiment with fifteen volunteers (all friends). The purpose of the experiment was to establish that textures on objects are used for depth perception. I ended up not using the results in this paper, but I thought I'd record the experiment here.

Get six standard coffee cups: 2 small, 2 medium, 2 large. For 3 cups: cover them in white paper, paint 3 dots on one side, and paint the same 3 dots on the reverse side plus 2 additional dots.

Give each participant a sheet of paper to note his answer on. Make sure they do NOT announce their responses, but to write them down instead.

Hand subjects the 3 unpainted coffee cups to give them an idea of how big they are. Now go 15-20 ft in front of them. Hold a small cup, with 3 dots. Move forward and backward +3 and -3 ft. How many people were able to correctly identify the size of the cup? (Since, if the actual dimensions of the cup are known, estimating these dimensions at a distance is precisely equivalent to estimating distance to the cup). Repeat for medium and large cups, in random order.

(The cups should be handed to the participants as opposed to shown to them: This is so that they do not develop a conception of how big the cups appear from a distance).

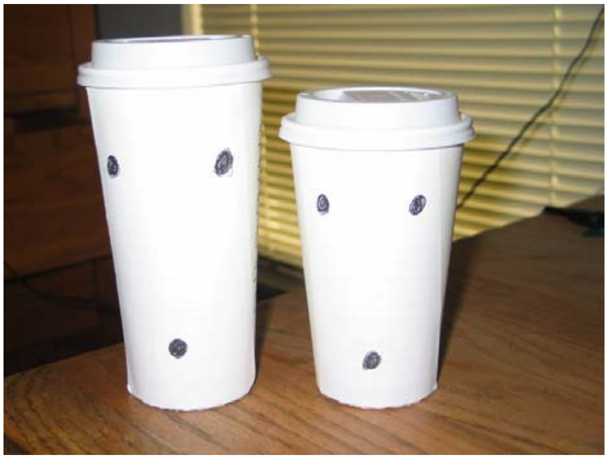

Figure A1. Featureless cups with 3 dots

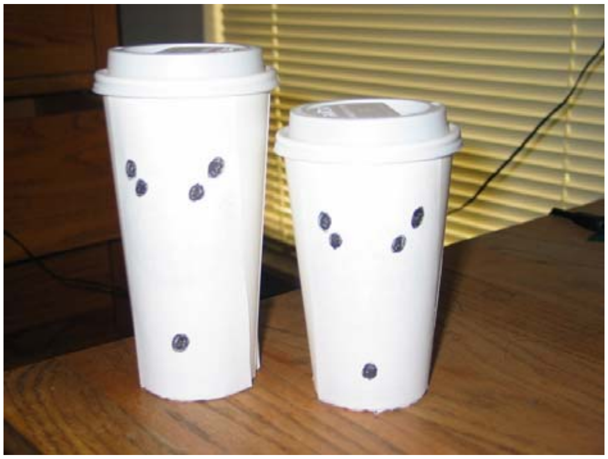

Figure A2. Featureless cups with 5 dots

The idea was that drawing the 2 additional dots so close to the previous three would add very little perceivable 3-D information to the cups, and so if there was any other factor in determining distance to the object (some kind of optical illusion) that was dependent on texture, we would see its effects in adding these two dots. If depth perception did not change, then we could try the same experiment again with dots more widely separated, and this should theoretically improve depth perception. As expected, participants performed equally well in cases with 3 dots and 5 dots. Unfortunately, there was sufficient disparity in the sizes of the small, medium, and large cups, that all subjects identified the sizes of the cups correctly every time, so it is not clear that the dots had anything to do with it. The experiment will need to be repeated with cups that are more closely matched in size.