

VISUAL RECOGNITION BASED ON INTEGRATED SENSORY MODALITIES

REZA SHAHBAZI

Cornell University, Department of Psychology

INTRODUCTION

The field of computer vision has been in constant progress for the last couple of decades, with new ideas and techniques some of which are no doubt major steps forward. In particular integration of machine learning techniques into the field shifted the machine vision researchers' views to a more statistics based solution to the problem of vision. Nevertheless, even the most complicated vision algorithms still perform inferior compared to human vision. Part of this issue may be due to the fact that the majority of computer vision techniques rely on merely visual content of the stimuli. This is in contrast to human vision in that at any instance of a visual task, all other sources of information (e.g. auditory input, tactile input, semantic knowledge, etc.) are integrated together to make in inference about the nature of the object/ scene in question [1,2]. Based on this inspiration, in this paper we attempt to simulate recognition of stimuli based not only on their visual content, but also on other sources of information available to the system. In particular, we are interested in comparing the performance, in terms of accuracy of visual recognition, of two similar systems where one relies only on the visual aspect of the data, whereas the other one also makes use of the auditory information.

The proposed approach in this paper is not entirely new. Other researchers have formulated a similar approach under the title “Cue Integration” (for example [3] and [4]). Our approach differs from previous ones in two ways.

First, most previous applications of cue integration technique have a discriminative nature. On the other hand, the human visual system has to deal with noise and ambiguity. This requires that the brain employs top-down processing in addition to bottom-up inference to perform the task of recognition with minimum error [5]. For this reason we take a generative approach in our application of the cue integration technique. Second, there exist instances of a generative application of cue integration method in the literature. However, they typically rely on different visual cues of the stimuli such as depth, edge distribution, color, etc. whereas, our simulation is based on both visual and auditory information of the data.

SIMULATION

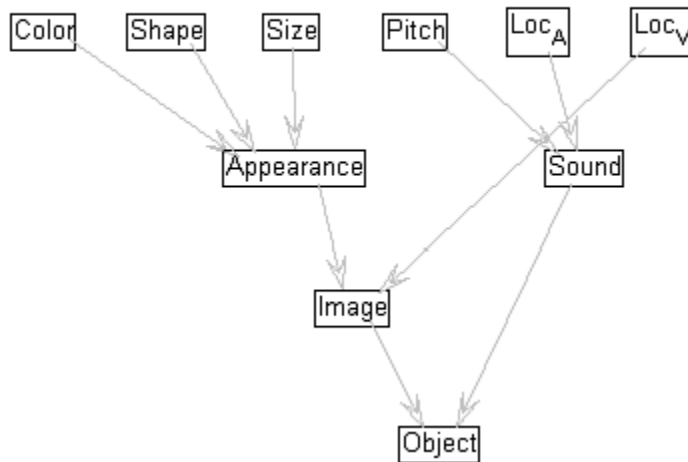
The first phase of this project involves comparing the performance of a generative classifier when trained on visual stimuli alone, versus when trained on both visual and auditory stimuli. This means is achieved by training two similar classifiers on the same set of data and then comparing their performance through ROC curves. Details are as follows.

We construct a probabilistic graphical model in the form of a directed acyclic graph as shown in figure 1. All nodes have discrete probability distributions, drawn at random from a Dirichlet distribution. In the next step we construct two more networks. One has an identical structure to our original network, while in the other one only those random variables that represent the visual part of the system are present. We then draw 1000 samples from our original network. We use 500 of these samples to train our two new networks, and another 500 for testing. When training the visual-only network, the fields of the sample set corresponding to auditory values are removed.

For training purposes, the conditional probability distributions of the nodes are first initialized at random, and assumed to have a Dirichlet prior (i.e. pseudo-count). The training samples are then used to learn the maximum likelihood parameters of each random variable. Training is performed in batch form.

ROC curves for the 10 possible outcomes of the top node are plotted as the true positive rate versus the false positive rate in Figure 2.

Figure 1



EXPERIMENT

With the results of the simulation supporting the general hypothesis (i.e. auditory information can improve visual recognition), we then proceed to conduct a similar experiment using real world data. To keep the experiment simple, we decided to use low level features of the stimuli. In particular, color histograms of the images were used as the visual feature of the stimuli, and the Fourier transform of the audio clips were used as the auditory feature.

Ten different bird species were selected (see below) as the target of classification. The objective of the experiment was to correctly classify an image into the correct bird species category using either visual features alone, or both visual and auditory features.

We downloaded video clips of ten different bird species from the web. Between twenty and hundred frames were extracted from each clip for a total of 524 images. Figure 3 shows samples of the ten bird species. Also between three and eight sound clips with a duration of one second were extracted from each clip for a total of 43 auditory features. Sound clips were sampled from those parts of the clip that contained the birds' voices. Color histograms were computed for each image sample and used as the input visual feature. Similarly, frequency histograms for the entire duration (1 sec) of each audio clip were computed and used as the input auditory features. Two third of the samples were used for training and one third were used for testing. Figure 4 shows the ROC curves for the classifier.

Figure 2

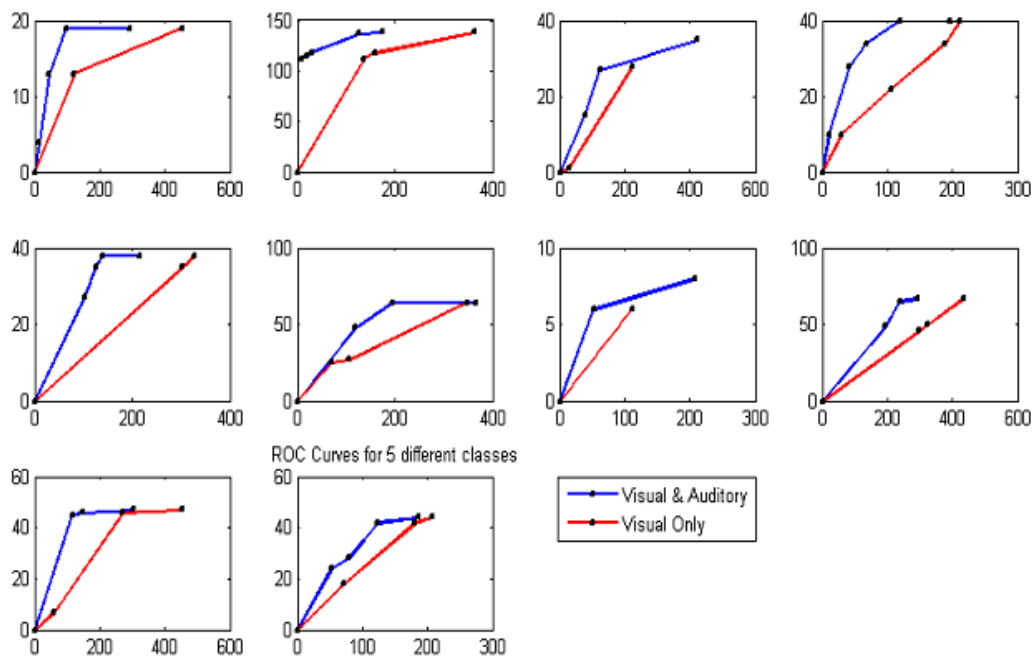


Figure 3

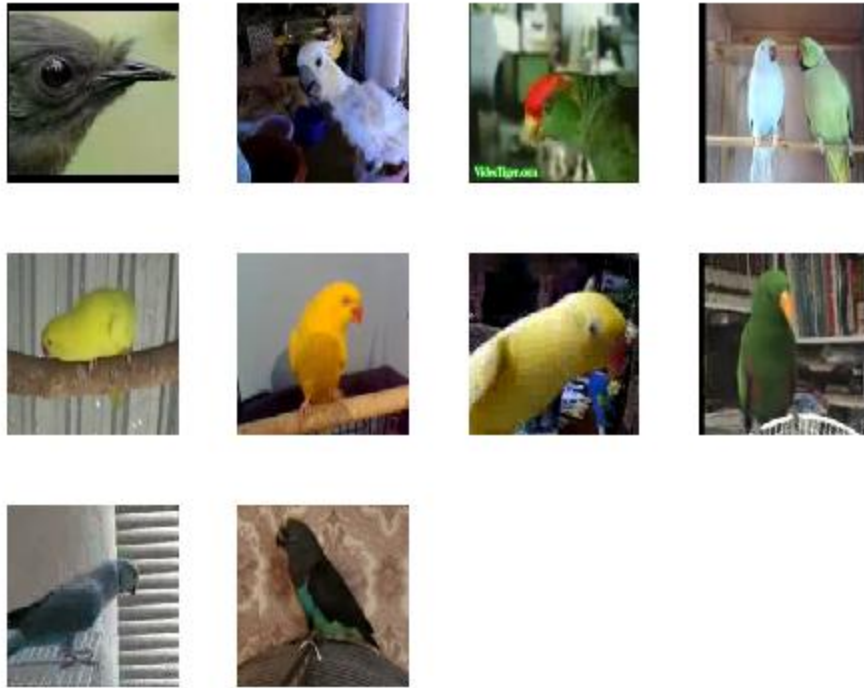
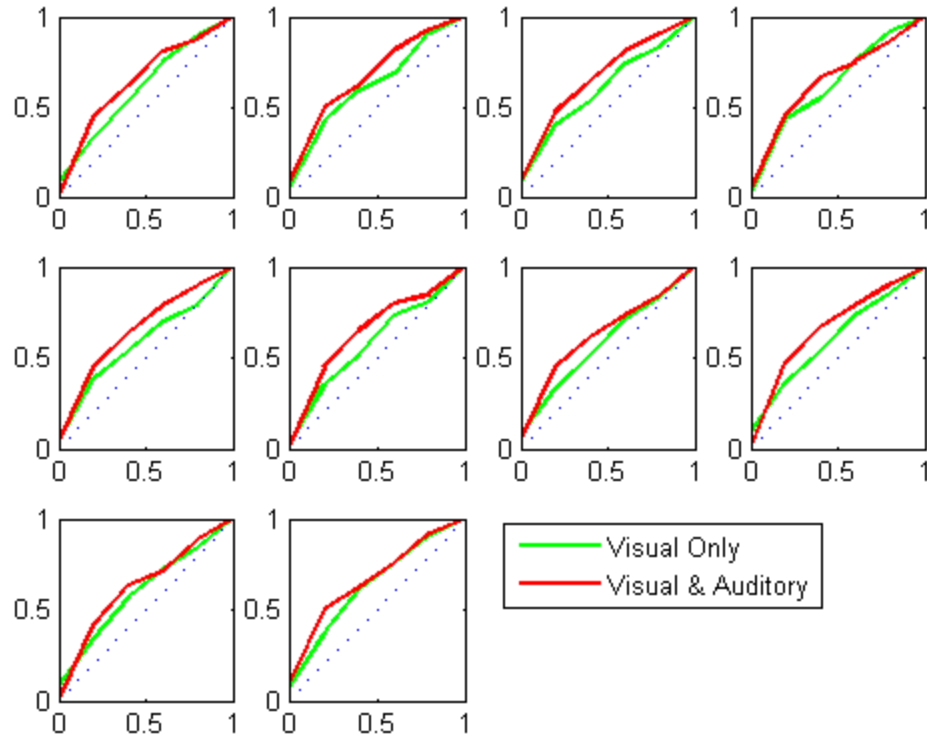


Figure 4



DISCUSSION

We have provided evidence supporting the hypothesis that the task of visual recognition can be aided by gathering the other sources of information available to the system which are not visual in nature.

As described earlier, in our experiment we relied only on low level statistics of the data. The primary reason for this choice was to keep the experiment reasonably simple to be completed within the designated time of submission to CS 6780. Nevertheless, this project is still a work in progress. The next step is to employ a hierarchical structure to extract higher level features of the stimuli. Our plan is to start with Gabor filters at several orientations and spatial

frequencies. We would then use mixture modeling (most likely with Gaussian components) to build higher level features in an unsupervised way.

BIBLIOGRAPHY

- [1] R. A. Jacobs. Optimal integration of texture and motion cues to depth. *Vision Res*, 39(21):3621–9, 1999.
- [2] R. J. van Beers, A. C. Sittig, and J. J. Gon. Integration of proprioceptive and visual position-information: An experimentally supported model. *J Neurophysiol*, 81(3):1355–64, 1999.
- [3] S Deneve, PE Latham, A Pouget -Efficient computation and cue integration with noisy population codes, *nature neuroscience*, 2001
- [4] M Spengler, B Schiele - Towards robust multi-cue integration for visual tracking, *Machine Vision and Applications*, 2003 – Springer
- [5] D., Kersten, *Object Perception: Generative Image Models and Bayesian Inference*, *Biologically motivated computer vision*, 2010 - Springer