Physica A 430 (2015) 1–10

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Classification error analysis in stereo vision

Eitan Gross

Department of Physics, 226 Physics Building, University of Arkansas, Fayetteville, AR 72701, USA

HIGHLIGHTS

- Depth perception model is presented in which mutual information in stereo vision can be analyzed using the theory of large deviations.
- For Gaussian neurons and a discrete input variable our model predicts that the mutual information saturates exponentially with N.
- For large N, the mutual information saturates exponentially with a rate determined by the Chernoff distance.

ARTICLE INFO

Article history:
Received 20 July 2014
Available online 3 March 2015

Keywords:
Chernoff distance
Large deviations theory
Mutual information
Random-dot-matrix
Saddle point approximation
Stereopsis

ABSTRACT

Depth perception in humans is obtained by comparing images generated by the two eyes to each other. Given the highly stochastic nature of neurons in the brain, this comparison requires maximizing the mutual information (MI) between the neuronal responses in the two eyes by distributing the coding information across a large number of neurons. Unfortunately, MI is not an extensive quantity, making it very difficult to predict how the accuracy of depth perception will vary with the number of neurons (N) in each eye. To address this question we present a two-arm, distributed decentralized sensors detection model. We demonstrate how the system can extract depth information from a pair of *discrete* valued stimuli represented here by a pair of random dot-matrix stereograms. Using the theory of large deviations we calculated the rate at which the global average error probability of our detector; and the MI between the two arms' output, vary with N. We found that MI saturates exponentially with N at a rate which decays as $1/N$. The rate function approached the Chernoff distance between the two probability distributions asymptotically. Our results may have implications in computer stereo vision that uses Hebbian-based algorithms for terrestrial navigation.

1. Introduction

In humans, acute depth perception of an unfamiliar scene is attained via binocular disparity by comparing the stereo images generated by the two eyes to each other [1,2]. However, given the highly stochastic nature of neuronal activity, it is not clear how our brain extracts acute depth information from the small difference between two nearly identical images obtained at slightly different angles of observation. One hypothesis [3] is that the brain is "searching" for features that are more stable than the noise by maximizing the variance of the sum of the neuronal response generated by the two eyes, divided by the variance of their difference. Making that assumption, Becker and Hinton [4] have designed an *unsupervised* neural network made of a two-arm detector, in which the two arms strive to produce outputs that agree with each other by maximizing the variance in their responses. This was implemented using the $\mathrm{I}_{\mathrm{max}}$ method [4] which maximized the variance of the *sum* of the two arms outputs divided by the variance of their *difference*. Assuming both the underlying signal and the noise have Gaussian distribution, this is equivalent to maximizing the mutual information (MI) between the outputs of the two arms. On

E-mail address: dreitangross@gmail.com.

http://dx.doi.org/10.1016/j.physa.2015.02.083

the premise that the accuracy of information presentation in the brain can be increased by distributing the information over a large number of neurons [5,6], one would expect the MI in the Becker–Hinton model to increased by increasing the number of nodes in their detector. MI is not an extensive quantity however, as it is bounded from above by the stimulus entropy; thus making it very difficult to predict how the accuracy of binocular disparity coding depends on the number of neurons (N) in the network. To address this issue, we present in the current paper a binary hypothesis detection system designed to extract depth information from a pair of discrete valued input stimuli represented by a pair of random dot stereograms (RDS). This is achieved by comparing, via the process of cross-correlation, the dot intensity (gray-level) in corresponding, or cross-matched, locations of the two (left and right) images. Indeed, these two processes (cross-correlation and cross-matching), have been implicated as the two types of neural representations of binocular disparity [7,8]. These representations are characterized by disparity tuning curves of the neurons involved with perceptual decisions. In correlation-based representation, the amplitude and sign of the disparity tuning curve follow the cross-correlation between images projected to the left and right eyes; while in match-based representation, the matched features between the images from the two eyes determine the amplitude of the tuning function. Neurons involved in the two representations have contrasting tuning functions in response to anti-correlated RDSs. In anti-correlated RDSs, the luminance contrast of the dots is reversed between the two eyes: white dots are replaced with black dots, and black dots are replaced with white dots against a gray background. The anti-correlation inverts the sign of cross-correlation and eliminates the matched features between the two eyes. Accordingly, the correlation based representation has inverted tuning curves for anti-correlated RDSs relative to those for correlated RDSs [7,9], while the match-based representation has flat (zero or decreased amplitude) curves [10,11]. The correlation-based and match-based representations refer to those distinctive sets of tuning curves, but they do not refer to underlying neuronal mechanisms.

The RDS-based system used in the current study reports a “hit” (or “1”) if it detects a match and “miss” -0- for a mismatch. Each of the two arms in our detector model contains N identical and independently distributed sensors, arranged in a parallel network architecture, that make local decisions based on their own measurements of the scene. Decisions made at the local level are transmitted *via* a noisy communication channel to a fusion center at the output unit of the channel which then makes a binary decision. The model enabled us to use concepts from the theory of large deviations, namely the saddle point approximation, to calculate the probability of making an error in the detection. From the error probability we calculated the rate at which MI between the two output units increases with N. We found that for a discrete valued input, MI increased exponentially at a rate significantly higher than the Chernoff distance. The rate function decayed inversely with N and approached the Chernoff distance asymptotically at large N. Model predictions were in good agreement with computer simulations based on the Becker–Hinton’s $\mathrm{I}_{\mathrm{max}}$ algorithm. Our model can be implemented in the design of Hebbian-based machinery that uses stereo vision for terrestrial navigation.

2. Model

We adopted the detection model for distributed sensors of Tenney and Snadell [12]. The model has been used in a variety of applications including surveillance [13], microwave imaging of earth surface [14] and to evaluate the receiver operating characteristics of a system of decentralized sensors [15]. Our Bayesian detector model minimizes the average global probability of error using the log-likelihood of the measurements as described in Ref. [15]. A log-likelihood based decision is not only more intuitively appealing than a one based on an ad hoc definition of error, but also makes more accurate decisions.

We start by considering the binary decision problem of deciding whether an input signal, representing a single spot in the left or right image of a RDS, is filled (H_1) or empty (H_0). Our network consists of two arms A and B, each containing N parallel input sensors, $\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_N$ (Fig. 1) and each looking at one of the two stereo images of the RDS. The network acquires N independent measurements, one per sensor, $y_n,\ n = 1, 2, \ldots, N$, makes a local b-bit decision u_n, i.e., quantizes the pixel’s gray level measured by the sensor into b-bits, delivers all N decisions $u_n,\ n = 1, 2, \ldots, N$, to a single output unit which makes the final determination H. The quantization performed locally in each sensor follows local decision rules $\gamma_n,\ n = 1, 2, \ldots, N$, in which the continuous observation space R representing the gray level intensity is being mapped into the discrete classification space U, i.e., $\gamma_n : R \to U$, where $U = \{1, 2, \ldots, M\}$ and M is the number of quantization levels. The quantization levels used here are analogous to the rate code used by neurons in our brain to encode and represent sensory information [16]. Upon receiving the N local decisions, the output unit in each arm fuses them together according to a fusion rule $\gamma_o : U^N \to \{0, 1\}$ to reach the final decision $\hat{H}$. The local decisions are based on the likelihood ratio (LR):

$$LR_{\boldsymbol{u}}\left(\boldsymbol{u}\right) = \frac{\Pr(\boldsymbol{u}|H_1)}{\Pr(\boldsymbol{u}|H_0)} \tag{1}$$

where $\boldsymbol{u} = (u_1, \ldots, u_N)$ is the vector of all quantized local decisions received at the output unit. Our Bayesian detector minimizes the global average probability of error:

$$P_e = \pi_0 P_0 + \pi_1 P_1. \tag{2}$$

P_j in Eq. (2) is the probability of error $\Pr\left(\hat{H} \neq H|H_j\right)$ and $\pi_j = \Pr(H_j),\ j = 0, 1$, are the prior probabilities. Here, we are interested in evaluating the hit and miss error probabilities, P_1 and P_0, given by:

$$P_0 = \Pr\left(u_0 = 1|H_0\right) \quad \text{and} \quad P_1 = \Pr\left(u_0 = 0|H_1\right) \tag{3}$$