Towards an information theoretic framework for object recognition

M. Brandon Westover and Joseph A. O'Sullivan
Dept. of Physics and Dept. of Electrical and Systems Engineering
Washington University in Saint Louis
Saint Louis, MO 63130
e-mail: jao@ee.wustl.edu, westovem@medicine.wustl.edu

Abstract — **We develop a generic model of object recognition systems from an information theoretic viewpoint. We derive inner and outer bounds on the conditions under which asymptotically error-free recognition is possible, in terms of a given statistical model for the environment and observation process, the number of patterns to be recognized, and the rates of the codes used to represent sensory data and memorized patterns.**

I. Introduction

A central goal of systems neuroscience is to discover how animal brains recognize the objects in their environments. Typical studies of animal object recognition are designed to elicit informative behavioral errors which can be analyzed to infer the computational strategies used by the animal. A difficulty often neglected in such experiments is the need to dissect performance limitations internal to the animal (e.g. suboptimal neural algorithms; implementational constraints) from those inherent to the problem (e.g. ambiguous data) [1]. Engineers face similar challenges in designing and analyzing artificial pattern recognition systems. Because recognition problems in real world environments tend to be ill-posed or intractable (e.g. due to high dimensional, complex, ambiguous data), a trial-and-error approach is often employed to discover practical recognition algorithms. How are performance limitations due to algorithmic and implementational deficiencies to be disentangled from those arising from the nature of the recognition problem itself? An overarching theoretical framework is needed for discussing and analyzing object recognition systems.

We believe that information theoretic concepts will be key components of an eventual unified mathematical framework for recognition. As a step in this direction, we propose an idealized mathematical model of object recognition systems. We derive two theorems which give inner and outer bounds on the performance achievable by the system as a function of statistical models for the environment, the observable data, and the sensory and memory subcomponents. The information theoretic perspective provided by our model and analysis helps to organize several key information processing issues faced by recognition systems, and provides the beginnings of a theoretical framework for normalizing system performance by the information requirements intrinsic to the recognition problem.

[1]This work was supported by the Mathers Foundation and by the Office of Naval Research, Award N00014-03-1-0110.

II. A mathematical model for object recognition

A block diagram for our idealized recognition problem is shown in Figure 1. In the present section we describe the essential components of the system. In the section following, we investigate conditions necessary and sufficient for the *existence* of reliable pattern recognition systems; the details of specific solution strategies are outside the focus of this paper. While comments in the following subsections apply equally to biological and artificial systems, for convenience we phrase the discussion in terms of a hypothetical unspecified animal. We illustrate issues discussed using examples from visual recognition.

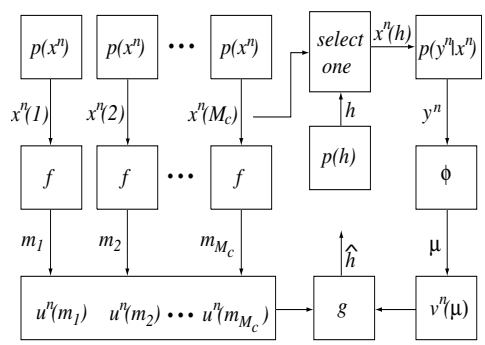

Fig. 1. Block diagram for a generic recognition problem.

A. Object source

The computational resources required to recognize objects depend on the complexity of an animal's environment. We let our animal's 'world' derive from a finite set of possible objects, $\mathcal{W}$. Of these, a subset $\mathcal{T} \subset \mathcal{W}$ constitutes our animal's specific ecological niche. We refer to $\mathcal{T}$ as the animal's *training set*: it contains the set of objects the animal must learn to recognize. We view the elements of $\mathcal{T}$ as functions $o(\cdot)$ (object models) parameterized by a (usually large) set of variables, $x^n = (x_1, \ldots, x_n)$, $x_i \in \mathcal{X}$, where $\mathcal{X}$ is a finite alphabet. In this paper, we let $o(\cdot)$ be generic, so that objects can be completely specified by giving the parameter vector alone; i.e. specifying $o(x^n)$ or simply x^n uniquely describes an object. In a visual recognition problem, the parameters might include viewing position and pose of the object; reflectance values at points on the object surface; or shape information. Alternatively, it may be convenient to define visual objects in terms of the luminance patterns they produce on the animal's retina, in which case x^n might

represent coefficients in a wavelet transform of the retinal image. When no confusion is possible, we usually refer to the objects in $\mathcal{T}$ using the notation $x^n \in \mathcal{T}$ rather than $o(x^n) \in \mathcal{T}$.

The creation of the animal's training set $\mathcal{T}$ is modeled as a random selection of some finite number M_c of vectors, $x^n(1), \ldots, x^n(M_c)$ from a distribution $p(x^n)$. We stipulate that, while the animal may inherit a genetically hard-wired disposition toward recognizing important environmental patterns and in this sense know $p(x^n)$, it cannot pre-know which specific objects $\{x^n(i)\}_{i=1}^{M_c} \in \mathcal{T}$ it will need to learn to recognize. Furthermore, the amount of environmental information available for training is generally far too much for the animal to fully internalize, both over the long term into memory, and via the sensory information available on any given occasion. These two considerations dictate that the animal must be equipped with memory and sensory representations that are both general purpose and efficient at deriving useful, compressed representations of the environment.

B. Memory

At some level, all object recognition strategies revolve around comparing sensory data with information stored in memory. In real systems, memories are always *approximate*, i.e. compressed, representations of the training data. There are several reasons for this. First, natural sensory signals are typically highly redundant, and can therefore withstand significant compression with little or no loss of vital information. Discarding irrelevant information may even actually facilitate later recall and pattern matching. Second, physical memories have limited precision and reliability. For example, in biological memory circuits, neurons may die, connections may degrade, and the recall process may be imperfect. Third, resource and energy costs associated with storing and maintaining memories favor compressed formats. Finally, the quantity of available raw data renders veridical memory impossible in practice.

We represent the learning module or *encoder* which derives memories from the (labeled) training data by a function $f(\cdot)$. In keeping with conventions of information theory, we view f as a mapping between the training vectors and a set of integers $f : \mathcal{X}^n \to \{1, \ldots, M_x\}$, where $M_x = 2^{nR_x}$, and R_x is the the *rate* of the code for the training data. These numbers index a set of 'physical' memories stored in our animal's brain, represented by vectors $u^n(i) \in \mathcal{U},\ i = 1, \ldots, M_x$, where $\mathcal{U}$ is an appropriate alphabet for storing memories (e.g., a repertoire of neural firing patterns).

C. Scene

For our purposes, testing of our animal takes place only *after* the training data have been learned. Recognition trials take place in the presence of a *scene*, consisting of one or more objects selected randomly from the training set. For simplicity, we will focus on the case in which only one object is present in any given test scene (although the actual sensory data may also include interfering signals, e.g. clutter; see the discussion of the sensor model below). In this case, a single object x_h^n is selected for inclusion in the scene by drawing one of the M_c training set indices h at random from a distribution $p(h)$. The recognition system must infer the object identity from an observation of the scene. That is, the system produces a guess $\hat{h}$, with the goal that $\hat{h} = h$.

D. Sensor

The vector x_h^n selected for recognition represents the ground truth reality, containing full information about the object. By contrast, the data observable by the animal y^n is generally incomplete, ambiguous, and distorted. In other words, x_h^n represents a collection of 'hidden variables' underlying the observed data, y^n. Observations will be modeled as a sequence of symbols y^n produced by a stochastic mapping $p(y^n|x^n)$, which we call the sensor model. We use the term 'sensor' quite broadly, to reflect what may be a very complex transformation between what is 'out there' and the observable data. In vision, for example, what we call the sensing process may include effects like the introduction of background clutter and occlusion to the image; blurring of light by the eye's lens; and neural noise and distortion introduced via early stages of visual processing.

E. Sensory encoding

We interpose a compression stage between the sensory output and the recognition module. While the details of encoding will typically be different for memory formation and sensory encoding, the motivations for compressing the data are similar to those mentioned earlier: Compression is inevitable given the unwieldy amount of sensory data available and the limited-precision representational capabilities of brains and machines; may be helpful in making relevant but initially latent information more explicit; and is desirable as a means of economically utilizing neural resources.

Here, the term 'encoding' encompasses any operations on the sensory output which reduce the total amount of information present in the original signal. In biological vision, sensory coding may include recoding of the retinal-level pixel based representation into a multiscale wavelet-like representation, as apparently occurs in simple- and complex-cell populations in monkey primary visual cortex; and dynamics of cortical recurrent and feedback loops which actively 'sparsify' the neural representation, enhancing the activities of neurons representing especially informative features of the sensory input while silencing others. We group these together because our present objective is not to study *how to* best process the data, but rather to determine simply *how much information* must be retained to solve a recognition problem reliably.

The sensory encoder is represented by a function $\phi(\cdot)$, which maps sensor outputs $y^n \in \mathcal{Y}^n$ into one of $M_y = 2^{nR_y}$ binary numbers, $\phi : \mathcal{Y}^n \to \{1, \ldots, M_y\}$. R_y is the *rate* of the sensory code. As in the learning module, these numbers index a set of codewords $v^n(i) \in \mathcal{V}^n,\ i = 1, \ldots, M_y$, where $\mathcal{Y}$ is the sensory alphabet. The codewords are the final physical representation of the sensory data available to the recognition module.

F. Recognition module

Recognition is accomplished by a module g, representing a statistical procedure for inferring the object class, subclass, etc., and attributes or state of the object in a scene. While in general there may be multiple objects in the scene, our present discussion is limited to the problem to inferring the identity of a single object h. We envision effective recognition procedures as variants of the Bayesian inference algorithms advocated by the Pattern Theory school of Grenander and colleagues [3]. However, in keeping with our information theoretic objectives, we leave details of the recognition algorithm unspecified in our discussion.

The recognition module has access to the memories $u^n(i)$ and encoded sensory data $v^n(j)$, and outputs a decision $\hat{h}$ to indicate which object x_h^n is presently in the scene. Thus, g represents a mapping $g(i,j) = \hat{h}$ from the index sets of the memory and sensory codebooks to the index set of the training data, $g : \{1,\ldots,M_x\} \times \{1,\ldots,M_y\} \rightarrow \{1,\ldots,M_c\}$.

III. Achievable rates for pattern recognition

We have analyzed a simple version of the system described in Section II to answer questions of the sort: To reliably recognize the objects in a given environment, (i.e., the animals in a jungle), how much information must an animal or machine store in memory and extract from data via its senses? More abstractly, given a model $p(x^n)$ of an environment and a model $p(y^n|x^n)$ of the observable data, what are the relationships consistent with reliable recognition performance between: (1) the complexity of the training data, quantified by the rate $R_c = \log(M_c)/n$; and (2) the rates of the memory representation $R_x = \log(M_x)/n$ and (3) the rate of the sensory representation $R_y = \log(M_y)/n$?

We present the results of our analysis in two theorems below, after stating simplifying assumptions used in our analysis, and defining reliable (a.k.a., *achievable*) recognition rates, probability of recognition error, and the rate region for achievable recognition. At present, we have succeeded in deriving inner and outer bounds on the achievable rates; current work is focusing on closing the remaining gap between the bounds. Some instructive 'sanity checks' follow the theorem statements, which help illuminate the theorems. Finally, we sketch proofs for the inner and outer bounds.

IV. Definitions and theorem

Our theorem will address the system described in Section II, diagrammed in figure 1. We analyze the special case in which the symbols produced by the pattern distribution $p(x^n)$ are independent and identically distributed, and the sensor model $p(y^n|x^n)$ is memoryless. That is, we assume $p(x^n) = \prod_i p(x_i)$, $p(y^n|x^n) = \prod_i p(y_i|x_i)$. We need the following definitions:

Definition 4.1 (Probability of error): The probability of misrecognition is given by

$$P_e^n = \frac{1}{M}\sum_{i=1}^{M} Pr\{\hat{h} \neq i | h = i\}.$$

Definition 4.2 (Achievable rates): A rate triple (R_c, R_x, R_y) is *achievable* if there exist sequences of $(2^{nR_c}, 2^{nR_x}, 2^{nR_y}, n)$ codes such that $P_e^n \rightarrow 0$ as $n \rightarrow \infty$.

Definition 4.3 (Achievable recognition rate region): For given environment and observation models $p(x^n)$ and $p(y^n|x^n)$, the corresponding *achievable rate region* is the set of achievable (R_c, R_x, R_y) triples.

Intuitively, an optimized system will use minimal resources to discriminate a maximal number of patterns. That is, good systems will operate near the surface of the achievable rate region, where R_c is large and R_x and R_y are small. The achievable recognition region is given by the following theorem.

Theorem 4.4 (Inner bound on the achievable rate region): Suppose there exist random variables U and V defined on finite alphabets $\mathcal{U}$ and $\mathcal{V}$, satisfying a length-four Markov chain constraint $U - X - Y - V$. Then all rates (R_c, R_x, R_y) satisfying

$$\begin{aligned} R_x &\geq I(U;X) \\ R_y &\geq I(V;Y) \\ R_c &\leq I(U;V) - I(U;V|X,Y). \end{aligned}$$

are achievable.

Theorem 4.5 (Outer bound on the achievable rate region): Conversely, if the rate triple (R_c, R_x, R_y) is achievable, then there exist random variables U and V defined on finite alphabets $\mathcal{U}$ and $\mathcal{V}$, satisfying two length-three Markov chains $U - X - Y$ and $X - Y - V$, such that the rates satisfy the inequalities

$$\begin{aligned} R_x &\geq I(U;X) \\ R_y &\geq I(V;Y) \\ R_c &\leq I(U;V) - I(U;V|X,Y). \end{aligned}$$

Remark 4.6: Note that in the last inequality in of inner bound, the last term $I(U;V|X,Y) = 0$, due to the length-four Markov chain $U - X - Y - V$. The term is nevertheless included to highlight the similarity between the inner and outer bounds. The essential difference separating the inner and outer bounds is the stricter Markov condition on the inner bound. Note the outer bound region is indeed larger than the inner bound region, as the stronger length-four Markov condition implies the two length-three constraints.

Remark 4.7: Note that if either $X = U$ or $Y = V$, or both, then the inner and outer bounds are identical. In such cases, one or both of the two short Markov conditions of the outer bound trivially satisfied, e.g. $X - X - Y$ or $X - Y - Y$. Equivalently, the Markov condition of the inner bound degenerates into a single length-3 Markov chain. Similar comments apply if U and V are any deterministic functions of X and Y.

Remark 4.8: We remark finally that there are are strong parallels between the gap between inner and outer bounds in our current problem and the known bounds on the achievable rate region for the the two-terminal distributed source coding problem, derived by Berger in [5].

V. Sanity checks

To build some intuition regarding this theorem, it is instructive to examine a few extreme scenarios which have intuitive qualitative explanations, and for which the inner

and outer bounds are *equal*; see Remark 4.7. That is, in the following special cases, the theorems above precisely define the achievable recognition rates. We examine the effects of varying the sensory and memory capabilities of the system on the number of objects the system can learn to recognize by studying the inequality $R_c \leq I(U;V)$.

Unlimited senses and memory. First, consider a system for which the memory and sensory representations are not limiting factors in recognition performance. In this case, we can effectively treat the memories u^n and sensory representations v^n as if they were veridical; i.e. for the random variables U and V in the theorem we can put $U = X$ and $V = Y$. The theorem constraints then become $R_x \geq I(X;X) = H(X),\ R_y \geq I(Y;Y) = H(Y)$, and

$$R_c \leq I(X;Y) = H(X) - H(X|Y). \tag{V.1}$$

This result indicates that, in the absence of compression, the recognition problem is the standard communication problem: Transmit one of $M_c = 2^{nR_c}$ possible messages (objects) to a receiver (the recognition module). In this case, the objects can be thought of as codewords which are *directly* stored for *direct* comparison with the sensory uncompressed data. This is the setup of Shannon's channel coding theorem, which gives the rates at which reliable communication is possible as those below the mutual information between the source (equivalent to the memory here) and the received signals, $I(X;Y)$. This is exactly the condition expressed by (V.1). The condition translates into an upper bound on the number of objects the system may be trained to recognize through the definition $M_c = 2^{nR_c}$.

The effects of distortion and noise in the observation process are also cleanly isolated in this situation: An unreliable channel has a larger $H(X|Y)$, imposing a tighter restriction on the number of recognizable objects. In the limit that observations are completely uninformative, $I(X;Y) = 0$, and $R_c = 0$, so that the system is useless.

Unlimited memory, limited senses. Next, suppose that memory is effectively unlimited, so that we can put $U = X$, but sensory data may be compressed. In this case, we can readily rewrite the condition on R_c as

$$R_c \leq I(X;Y) - I(X;Y|V). \tag{V.2}$$

We check the extreme cases: If we let V contain full information about the sensory data Y, then $I(X;Y|V) = H(Y|V) - H(Y|X,V) = 0$, and we recover the case discussed above. On the other hand, if V is uninformative, then $I(X;Y|V) = I(X;Y)$, and we get a useless system with $R_c = 0$.

For intermediate cases equation (V.2) is also qualitatively very reasonable. If V is partially informative, then the effect of V is to degrade the achievable performance of the system below that possible for the 'perfect senses' case by subtracting $I(X;Y|V)$.

Limited memory, unlimited senses. In the case of possibly poor memory but effectively unlimited sensory representational abilities, we get an expression symmetric with the previous case:

$$R_c \leq I(X;Y) - I(X;Y|U). \tag{V.3}$$

As before, if we allow the memory to be perfect ($U = X$), we get $I(X;Y|U) = I(X;Y|X) = 0$, recovering the channel coding constraint $R_c \leq I(X;Y)$; assuming useless memories yields $R_c \leq I(X;Y) - I(X;Y) = 0$; and intermediate cases place the system between these extremes.

VI. Sketch of Inner and Outer Bound Proofs

We now proceed to sketch the proofs. The proofs rely on standard techniques from information theory (c.f. [4]), with which we assume the reader is familiar.

A. Inner bound (Achievability)

For achievability, we show that for any $\mathcal{U}$, $\mathcal{V}$, $p(u|x)$, $p(v|y)$ and rates satisfying (IV.1), there exist a sequence of codes and recognition algorithms such that the probability of recognition error goes to zero as $n \to \infty$.

Our proof employs random coding and strong joint typicality arguments. Codebooks of 2^{nR_x} and 2^{nR_y} length-n codewords consisting of independent symbols are drawn independently from $p(u)$ and $p(v)$, and made available to the recognition algorithm. Memories and observations are encoded by giving the index of a jointly typical codeword, m and μ. Decoding is done by searching for a vector in the memory codebook $U^n(m)$ jointly typical with the sensory representation of μ, $V^n(\mu)$. The recognized object is declared as $\hat{h} = m$.

Without loss of generality, assume $h = 1$. The potential error events and reasons their probabilities vanish for large n are as follows. E_1: Pattern x_1^n is not represented in memory. Choosing $R_x \geq I(U;X)$ ensures that there are enough codewords to 'cover' the the training set, so $Pr(E_1) \to 0$. E_2: The observation y^n is unencodable. As for E_1, choosing $R_y \geq I(V;Y)$ ensures that $Pr(E_2) \to 0$. E_3: A pattern other than x_1^n is jointly typical with the same memory vector U^n. This possibility is made improbable via the source coding theorem if there are not 'too many' patterns to be recognized, i.e. if $R_c \leq I(U;X)$, which is guaranteed by choosing $R_c \leq I(U;V)$, since $U \leftrightarrow X \leftrightarrow Y \leftrightarrow V$ form a Markov chain. E_4: The encoded observation $V^n(1)$ is not strongly jointly typical with the appropriate memory $U^n(1)$. The argument for $Pr(E_4) \to 0$ is more subtle, but follows from the Markov observation and applying Berger's well-known Markov Lemma [5]. E_5: The encoded observation $V^n(1)$ matches a memory other than $U^n(1)$. For the same reasons as for E_3, choosing $R_c \leq I(U;V)$ ensures that $Pr(E_5) \to 0$.

B. Outer bound (Converse)

To prove the outer bound (converse), we must show that for any sequence of codes and recognition algorithms for which $P_e^n \to 0$ as $n \to 0$, there exist a pair of random variables U, V consistent with the Markov conditions $U - X - Y$ and $X - Y - V$, such that the inequalities in (IV.1) hold.

Begin by defining auxiliary random variables $U_i = (m, X^{i-1})$ and $V_i = (\mu, Y^{i-1})$. These satisfy the Markov requirements. It is also straightforward to show that $nR_x \geq \sum_{i=1}^n I(X_i;U_i) = \sum_i I(X_Q;U_Q|Q = i) = nI(X;U)$,

where the last step follows from defining $U = (U_Q, Q)$ and standard time sharing arguments. Similarly $nR_y \geq I(Y;V)$, with $V = (V_Q, Q)$.

The final part of the proof is to show, for the same choices of U and V given above, that $nR_c \leq nI(U;V) - I(U;V|X,Y) + \epsilon$. A few straightforward manipulations and application of Fano's inequality give that $nR_c \leq I(m;\mu)$. A more difficult step (see next paragraph) establishes $I(m;\mu) \leq \sum_{i=1}^n I(U_i;V_i) - I(U_i;V_i|X_i,Y_i)$. The remaining steps to convert $\sum_{i=1}^n I(U_i;V_i) - I(U_i;V_i|X_i,Y_i)$ into $I(U;V) - I(U;V|X,Y) + \epsilon$ are once again routine time sharing arguments.

Proving the crucial intermediate step required generalizing a Lemma originating in [6] and recently reintroduced as Lemma 1 in [7] in the context of channel and source coding problems with state information.

Lemma 6.1: For any five random variables $A, \alpha, B, \beta, \gamma$,

$$\begin{aligned} I(A;\alpha) - I(B;\beta) \\ &= I(A;\alpha,\gamma) - I(A,\alpha;\gamma) + I(\alpha;\beta,\gamma) \\ &\quad - I(B;\beta,\gamma) + I(B,\beta;\gamma) - I(\beta;\alpha,\gamma). \end{aligned}$$

The proof follows immediately by applying basic information theoretic identities to the right hand side of the expression. Surprisingly, this simple result generalizes what was easily the most opaque part of the converse in [6].

To apply the Lemma, we make the substitutions $B = 0, \beta = 0, A = X_i, \alpha = X^{i-1}, \gamma = m$, and sum the resulting expression over $i = 2 : n$ to obtain (after some cancellations) $0 = \sum_{i=2}^n I(X_i;X^{i-1}) = \sum_{i=1}^n I(X_i;U_i) - I(X^n;m)$; the LHS is zero because X_i and X^{i-1} are independent. Making two similar substitutions and summing the results also gives $0 = \sum_{i=1}^n [I(X_i;U_i) + I(Y_i;V_i) - I(X_i,Y_i;U_i;V_i)] - [I(X^n;m) + I(Y^n;\mu) - I(X^n,Y^n;m,\mu)]$. Next, noting that we have the Markov chain $m \leftrightarrow (X^n,Y^n) \leftrightarrow \mu$, straightforward calculations show that the second bracketed term is equal to $I(m;\mu)$. Similar calculations on the bracketed terms in the first sum show that it is equal to the corresponding quantity $I(U_i;V_i) - I(U_i;V_i|X_i,V_i)$. Thus, $I(m;\mu) \leq \sum_{i=1}^n I(U_i;V_i) - I(U_i;V_i|X_i,V_i)$, which is the desired inequality.

VII. Conclusions

In this paper, we have described a mathematical model for object recognition problems from the perspective of information theory. Given stochastic models for the environment and data generating process, we derived inner and outer bounds on the conditions under which reliable recognition is possible in terms of three key quantities: the number of objects to be recognized, and the compression rates for storing data in memory and encoding sensory data.

Some caveats are in order. Our results apply to the case where the underlying object models having i.i.d. parameters, and a memoryless sensor model. While this case provides important insight, it be will of course be desirable in future work to relax these restrictions. Also, in this paper we have considered the scenario in which 'perfect' (i.e. error free) recognition is desired. More generally, an animal may only require partial or 'coarse' recognition; or the level of precision at which identification needs to be performed may vary dynamically. We are currently exploring information theoretic aspects of such 'coarse-to-fine' recognition strategies. Secondly, in our analysis we have assumed the availability of generative models $p(x^n)$ and $p(y^n|x^n)$ for the environment and observations. We are presently applying our results to analyze a model environment in which relevant objects are samples from statistical texture models (e.g. as in [8]). However, in more complicated environments, deriving adequate statistical models may be the most challenging aspect of the problem. Nevertheless, our results provide useful tools for analyzing such models as they are developed.

Finally, there remains a gap between the inner and outer bounds on the achievable recognition rates in the theory presented here. Current work is focused on closing this gap.

References

[1] P. Schrater, D. Kersten, "Vision, Psychophysics, and Bayes," In R. P. N. Rao, B. A. Olshausen, M. S. Lewicki (Ed.), Statistical Theories of the Brain. Cambridge, Massachusetts: MIT Press, 2002.

[2] C. E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.

[3] U. Grenander, "General Pattern Theory," Oxford University Press, 1993.

[4] T.M. Cover and J. Thomas, "Elements of Information Theory," Wiley, New York, 1991.

[5] T. Berger, "Multiterminal Source Coding," In G. Longo, editor, *The information theory approach to communications,* Springer Verlag, New York, 1977.

[6] S.I. Gel'fand and M.S. Pinsker, "Coding for channel with random parameters," *Problems of control and Information Theory,* vol. 9, no. 1, pp. 19-31, 1980.

[7] T.M. Cover and M. Chiang, "Duality between channel capacity and rate distortion with two sided state informatoin," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1629–1638, June 2002.

[8] S. C. Zhu, Y. N. Wu and D.B. Mumford, "FRAME: Filters, Random field And Maximum Entropy: — Towards a Unified Theory for Texture Modeling," International Journal of Computer Vision, 27(2) pp.1-20, March/April 1998.