FEATURE-BASED INFORMATION PROCESSING WITH SELECTIVE ATTENTION

Christopher J. Rozell, Ilan N. Goodman and Don. H Johnson

Department of Electrical and Computer Engineering
Rice University, Houston, TX 77251-1892

ABSTRACT

We present a simple but general model for feature-based information processing with selective attention. We model feature extraction as projections onto frames of subspaces, which accounts for redundancies in the representations of individual features as well as between features. To manage limited resources, we use feedback attentional signals to dynamically allocate system resources according to the observed events. In our model, attention maximizes the average information retained about all events weighted by their relative priorities. We illustrate the model with a simple system under a total bit constraint and discuss how the organization of the feature extraction affects the optimal bit allocation.

1. INTRODUCTION

Many of today's important information processing applications (sensor networks, for example) require intensive distributed processing of sensor data, but are highly constrained by the available resources such as energy and bandwidth. An analogous problem is faced by biological sensorineural systems, which use limited resources to process extremely complicated sensory information. Yet despite the constraints, sensorineural systems perform amazingly well even on very complicated tasks. Understanding the benefits of neural organizational principles will enable us to build more efficient distributed information processing systems.

Sensory systems have the common characteristic of first "ripping apart" a complex received signal into very basic component *features*, each representing specific information about the signal. These features are selectively combined to form more specific complex features in successive stages. In contrast to matched filters, which narrowly characterize the signal according to a single template, feature-based processing is flexible enough to describe new events as they arise. To accomplish this complicated signal decomposition and feature combining with limited resources, neural systems focus their processing on "important" events occurring in the sensory scene [1]. Through the process of *selective attention,* higher-level systems use feedback to control lower-level systems according to what seems most relevant and consistent with *a priori* models of perceptual features.

Email: {*crozell,igoodman,dhj*}*@rice.edu*

In order to assess the advantages of feature-based processing and selective attention, we must have a mathematical description of this information processing strategy. While much analysis remains to be done, we present here an initial description of a simple but very general model for feature extraction in a sensory system based on frame theory. In our model, attention is used to maximize the information preserved about important events *dynamically* as new events arise and old ones disappear. To demonstrate the usefulness of this model, we analyze a simple bit-allocation scheme in which features are encoded at different rates according to their relative importance.

2. FEATURE-BASED PROCESSING

Sensory processing in biological systems occurs in stages; front-ends form complete representations of signals in their operating range, and many parallel secondary areas each respond to a more stimulus-specific feature, such as color or motion. In every area, information is represented redundantly. Feature extraction systems must therefore resolve two distinct types of redundancy: (1) the redundancy *within* the representation of an individual feature, and (2) the redundancy *between* different features. We model this processing architecture with the mathematical tools of frame theory.

The sensing process is simply modeled as a collection of spatial and/or temporal filters corresponding to each sensor, which is equivalent to projecting a stimulus signal s onto a set of vectors in a high-dimensional space. Expanding a signal $s \in \mathcal{H}$ in a set of orthogonal basis vectors for $\mathcal{H}$ is a common operation in signal processing, $s = \sum_j k_j \phi_j$, with the coefficients $k_j = \langle s, \phi_j \rangle$ constituting the projections of the signal onto the associated basis vectors. However, most biological and man-made sensing systems use a collection of sensors that are not orthogonal. Therefore, the sensing operation is better described as a projection of signals onto a redundant collection of vectors (an overcomplete basis) known as a *frame* [2]. A collection of vectors $\{\phi_j\}$ is a frame for a Hilbert space $\mathcal{H}$ if there exist constants $0 < A \leq B < \infty$, known as frame bounds, so that for any function $s \in \mathcal{H}$ the Parseval relation is bounded,

$$A \|s\|^2 \leq \sum\nolimits_j |\langle s, \phi_j \rangle|^2 \leq B \|s\|^2 . \qquad (1)$$

When the frame vectors are normalized, A measures the

frame's minimum redundancy in covering the space $\mathcal{H}$.

In contrast to the sensing process, *feature extraction* occurs when some information is eliminated from the stimulus signal to isolate a more specific aspect of the signal. Specifically, feature extraction amounts to the projection of the signal onto a subspace of the signal space. Because different features represent some of the same aspects of the signal, feature subspaces may overlap in the signal space. Feature extraction stages are progressively more complex and more specific, meaning that higher-level features represent only a low-dimensional subspace of the original high-dimensional signal space. Mathematically, a feature is encoded by projections onto a (possibly redundant) set of vectors that span only a subspace of the input space $\mathcal{H}$.

The possibility of overlap between feature subspaces means that the total collection of features considered together may form a redundant representation for the whole signal space. The new theory "frames of subspaces" can be used to characterize these redundant feature subspaces [3, 4]. Here, a single feature space is represented by a collection of vectors contained in the rows of the matrix F_l. These vectors form a local frame for a feature subspace $W_l \subseteq \mathcal{H}$, with local frame bounds (A_l, B_l). The feature extraction operation is the projection onto W_l, given by $f_l = P_l s$. The frame coefficients representing a feature are corrupted by noise $c_l = F_l f_l + n_l$, which will model the limited fidelity imposed by resource constraints. The collection of (possibly overlapping) subspaces $\{W_l\}$ is called a *frame of subspaces* for $\mathcal{H}$ if there exist constants $0 < C \leq D < \infty$ such that for any $s \in \mathcal{H}$

$$C \|s\|^2 \leq \sum\nolimits_l \|P_l s\|^2 \leq D \|s\|^2. \qquad (2)$$

This Parseval's relation formalizes feature space overlap.

3. SELECTIVE ATTENTION

As the number of feature spaces in a system becomes larger, computing and transmitting all of the necessary coefficients overwhelms system resources. Sensorineural systems cope with this through the process of *selective attention*, wherein network resources are dynamically allocated to feature spaces according to what events are present in the scene. This allocation is mediated by top-down attention signals; feedback signals pass from high-level processes to the systems providing their inputs. In this way, attention focuses limited resources on what is important.

"Importance" in this setting has two components: the priority of an event relative to other events in the scene, and the relevance of each feature space to that event. Event priorities characterize how important each event is to the system's goals; for example, an event corresponding to the presence of an intruder would receive a high priority, whereas events corresponding to slowly changing ambient conditions would receive very low priority. The relevance of a feature space to an event corresponds to how much information that subspace preserves about the event, which depends on both the coherence of the event with the feature space and the redundancy of the underlying feature representation.

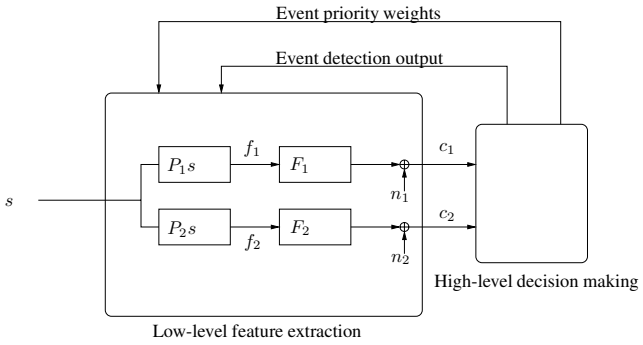

Fig. 1. Low level processing structures extract features from the signal by projecting onto overlapping subspaces with projection operator P_l. The outputs are represented by a redundant frame expansion for that subspace contained in the matrix F_l and the coefficients are represented with limited fidelity (represented by additive noise). Higher-level systems determine the events present and their relative priorities, and this feedback allows the feature extractors to adjust their fidelity to maximize the information about the important events.

In general, attention signals should allocate resources to each subspace in order to optimize a performance criterion related to the system's task. However, though tasks such as detection and estimation are likely, we seek to optimally assign resources without explicitly specifying the final goal of the system. Thus we require a general method to quantify the amount of information represented in the feature space output. We turn to a theory of *information processing* [5]. In this theory, quantifiable changes in a system's input are compared to the corresponding changes in its output. If the input signal is changed and there is (statistically) no change in the output, the system has not preserved any information about that event. However, if the output also changes to reflect the presence of the new event, then the output contains some information about that event. Quantifying this change measures how much information about that event is preserved by the system.

This method is best illustrated by an example: For simplicity, consider an input s generated from a Gaussian distribution centered on an event vector, $s \sim N\left(e_m, \sigma_s^2 I\right)$, and consider what happens when the input event is perturbed: $e_m \rightarrow e_m + \Delta$. Because the signals are stochastic processes, we measure change by calculating the distance between the probability laws that govern the original signal and its perturbation. There are several information theoretic distances that measure changes in probability distributions. We use the Kullback-Leibler (KL) distance [6] because of its relevance both to detection error decay rates and mean-squared estimation error [5]. For a Gaussian distribution with mean μ and covariance Γ, a change in mean Δ results in a KL distance of $D\left(\mu \| \mu + \Delta\right) = \Delta^t \Gamma^{-1} \Delta / 2$. In the stimulus signal space, the KL distance under the perturbational mean change

is therefore $D_s\left(e_m \| e_m + \Delta\right) = \|\Delta\|^2 / 2\sigma_s^2$.

Information can be lost in a feature space representation in two distinct ways: (1) the signal of interest is often not completely contained in the feature space, and (2) the redundant representation of each feature is imperfect due to limited communication resources. We want to determine how well the system as a whole preserves the information about an event. Consequently, we need to calculate the KL distance between e_m and $(e_m + \Delta)$ for the total collection of feature space encodings.

As illustrated in Figure 1, the stimulus signal is projected onto each feature subspace $f_l = P_l s = V_l V_l^t s$, where V_l is an orthonormal basis spanning the feature space $W_l \subseteq \mathcal{H}$. The extracted feature is represented by projecting it onto a collection of vectors (on the rows of the matrix F_l) that form a frame for W_l and adding noise, $c_l = F_l f_l + n_l = F_l P_l s + n_l$, where $n_l \sim N\left(0, \sigma_{n_l}^2 I\right)$. The noise term n_l here models the resource constraint, with $\sigma_{n_l}^2$ inversely related to the amount of resources allocated to the l^{th} feature. We need to calculate the joint KL distance under a mean change for the collection of feature outputs, $\overline{c} = \left[c_1^t, c_2^t, \ldots, c_L^t\right]^t$.

The KL distance across the joint collection of feature spaces under the mean change is given by $D_{\overline{c}}\left(e_m \| e_m + \Delta\right) = \frac{1}{2}\left(\overline{V}\Delta\right)^t K^{-1}\left(\overline{V}\Delta\right)$, where $\overline{V} = \left[V_1, V_2, \ldots, V_{L^t}\right]^t$, $K = K_s + K_n$ and

$$K_s = \sigma_s^2 \cdot \begin{bmatrix} I & V_1^t V_2 & \cdots & V_1^t V_L \\ V_2^t V_1 & I & \cdots & V_2^t V_L \\ \vdots & & \ddots & \\ V_L^t V_1 & & & I \end{bmatrix},$$

$$K_n = \text{diag}\left[I\frac{\sigma_{n_1}^2}{A_1}, \ldots, I\frac{\sigma_{n_L}^2}{A_L}\right].$$

In information processing theory, the important quantity is the "information transfer ratio," which is the ratio of the output to input distances: $\gamma_{s,\overline{c}}\left(e_m, e_m + \Delta\right) = \dfrac{D_{\overline{c}}\left(e_m \| e_m + \Delta\right)}{D_s\left(e_m \| e_m + \Delta\right)}$. The information transfer ratio measures what fraction of the input information is preserved in the collection of feature spaces. A fundamental result from information theory states that the information transfer ratio will always be between zero and one. A value of one represents perfect information transfer of the stimulus change; a value of zero represents total information loss. The goal of attention will be to assign system resources to feature spaces to maximize the information transfer about events according to their priorities. This task is complicated by the fact that events with different priorities may be represented to different degrees in the same subspaces.

To illustrate the attentional mechanism in a feature extraction system, we will consider a simple system in which network bandwidth is the scarce resource. Here, the combined data rate of all feature subspaces is constrained; each feature must be represented and communicated with a only a limited number of bits. The role of attention, then, is to dynamically allocate the available bits among the different subspaces to best represent the interesting events. Mathematically, given L feature subspaces and the system capacity of B_{total} bits, we allocate B_l bits to each feature subspace so that $\sum_l B_l = B_{\text{total}}$. We assume that there are M possible events, and that each event's priority is expressed by α_m $\left(\sum_m \alpha_m = 1\right)$. For our performance metric we use the weighted sum of information transfer ratios for all events $\{e_m\}$ that are determined to be present.[1] The optimal bit allocation is the one that maximizes the expression $\gamma = \sum_m \alpha_m \gamma_{s,\overline{c}}\left(e_m, e_m + \Delta\right)$ subject to the constraint $\sum_l B_l = B_{\text{total}}$. We assume the resulting noise variance in each feature space is given by $\sigma_{n_l}^2 = 2^{-2B_l}$, modeling the effect of uniform scalar quantization. This allocation reflects the relative priorities of the events, the coherence of each feature space with each event, and the robustness provided by the frame for each feature space.

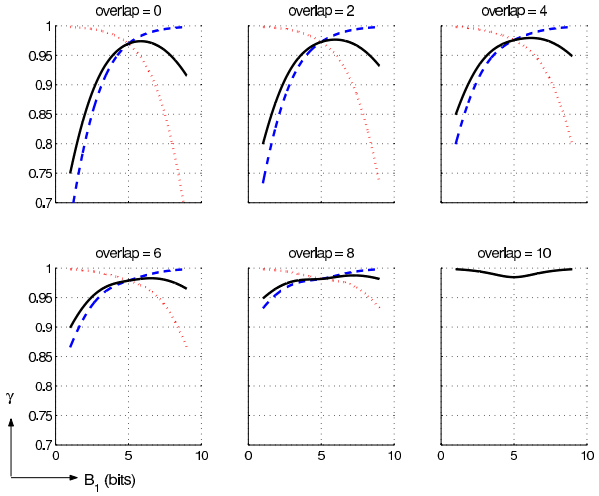

Fig. 2. Information transfer vs. bit allocation to one subspace for a two-feature system. Shown are the information transfer ratios for two events e_1 (dashed) and e_2 (dotted) with priorities $\alpha_1 = 3/4$ and $\alpha_2 = 1/4$. The solid curve is the priority-weighted sum γ. With increasing overlap between features, allocating more bits to W_1 maximizes γ.

To see how this bit allocation changes as a function of overlap between feature spaces, consider the following simple example. Assume the input space is $\mathcal{H} = \mathbb{R}^{10}$ and we have two subspaces W_1 and W_2 of equal dimension that form a frame of subspaces for the input space. For simplicity, each subspace is represented with an orthonormal basis $(A_1 = A_2 = 1)$. There are two events, $\{e_1, e_2\}$, each having unit norm, with priorities $\alpha_1 = 3/4$ and $\alpha_2 = 1/4$. Initially, we consider non-overlapping subspaces that each span 5 dimensions of the input space, such that $e_1 \in W_1$ and $e_2 \in W_2$.

[1] The presence of each event is determined in this example by an omniscient higher-level processor providing the attentional signals, but in a working system the salient events would be determined by a detection algorithm that would occasionally make mistakes.

Figure 2 shows the total information transfer ratio for each event, $\gamma_{s,\overline{e}}(e_m, e_m + \Delta)$, as well as the weighted sum γ as a function of the bit allocation for this system. The horizontal axis is B_1, the number of bits allocated to the space W_1 out of the total pool of $B_{\text{total}} = 10$ (meaning that $B_2 = 10 - B_1$). The bit allocation for the non-overlapping subspaces is plotted in the top left panel. Even though the information transfers for the two events are symmetric, e_1 has higher priority and the maximum γ is achieved by allocating more bits to W_1 than W_2.

In the remaining plots, the feature subspaces are enlarged by successively sharing basis vectors between the subspaces. For example, in the bottom left plot, each subspace covers eight dimensions of the input space, and they share six common dimensions. In this case, the optimal allocation assigns even more bits to W_1, since it now contains both the high-priority event e_1 as well as a significant part of e_2. In the extreme case when the subspaces overlap completely ($W_1 = W_2$), the optimal bit allocation assigns all the bits to one or the other subspace, since no new information is gained by including both. In fact, it is perhaps startling to note that dividing the bits equally between both subspaces results in the *worst* performance for this system. Qualitatively similar results were also seen with three feature subspaces.

In a real system, resource allocation must be dynamic; attentional signals need to adapt as new events arise and old ones disappear. Figure 3 illustrates this for the two-feature example discussed earlier. Here, there are four possible events in the space, indexed in order of increasing priority. The bottom plot indicates the presence or absence of each event at a given time. At each time step, the priority weights are adjusted to reflect only those events that are present. When the feature spaces have only two dimensions in common, the optimal bit allocation changes at almost every time step, reflecting the new priorities as the scene changes. As the features become more redundant, however, the bit allocation becomes more static; now, both subspaces convey significant information about all events, and the optimal solution tends to assign all the bits to a single feature subspace.

4. CONCLUSIONS

Using frame theory and frames of subspaces we have presented a very simple but general model of feature extraction. By drawing on the theory of information processing, we calculated in a general way the information present about an event in a collection of feature outputs. By maximizing the information transfer of the feature extractors weighted by the priorities of the events in the scene, we determined optimal allocation of resources to the individual feature spaces. Even the simple examples explored here show that blind allocation of resources can result in inferior performance. Attentional feedback allows the system to redistribute its resources adaptively to maintain optimal information transfer.

This paper describes an initial foray into a general theoretical framework for attentional processing in feature-based systems. In the examples we presented, bit allocations were only made on the basis of events that were known to be present. In a more advanced simulation a system will need to enforce lower bounds on the information present about each event so the appearance of new events isn't missed. While a total bit constraint was used here, the adaptation in each subspace could reflect other constraints such as communication power. The total information transfer expression does simplify somewhat and we believe that further analytic work will provide insight into achieving optimal resource allocation.

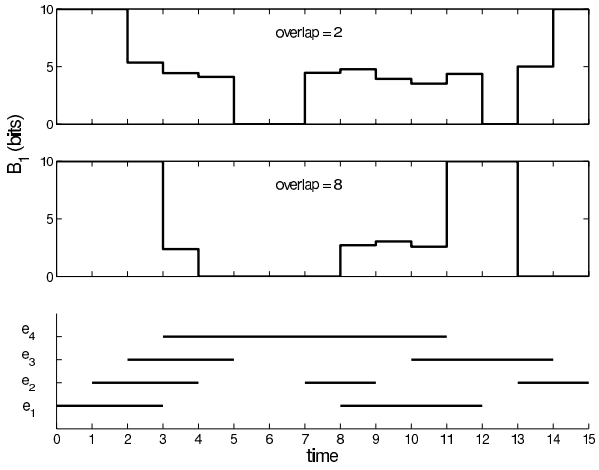

Fig. 3. Dynamic bit allocation in a two-subspace system. The top two plots depict B_1, the bit allocation to W_1 resulting from maximizing γ. At any time each of four events may be present. For the top plot, the two feature subspaces share two common dimensions, whereas for the middle plot they share eight dimensions. In the high-overlap case, bits are more often allocated to a single subspace. The bottom plot indicates the events present at each time step.

5. REFERENCES

[1] M.S. Gazzaniga, R.B. Ivry, and G.R. Mangun, *Cognitive Neuroscience*, W.W. Norton & Co., New York, second edition, 2002.

[2] O. Christensen, *An Introduction to Frames and Riesz Bases*, Birkhauser, Boston, MA, 2002.

[3] P. Casazza and G. Kutyniok, "Frames of subspaces," in *Wavelets, Frames and Operator Theory*. 2004, vol. 345, pp. 87–113, American Mathematical Society.

[4] C.J. Rozell and D.H. Johnson, "Analyzing the robustness of redundant population codes in sensory and feature extraction systems," *Neurocomputing*, 2006, In press.

[5] D.H. Johnson, C.M. Gruner, K. Baggerly, and C. Seshagiri, "Information-theoretic analysis of neural coding," *Journal of Computational Neuroscience*, vol. 10, pp. 47–69, 2001.

[6] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., New York, NY, 1991.