A Standard Model of Mind:
AAAI Technical Report FS-17-05

Elemental Cognitive Acts, and Their Architecture

Richard Granger, Eli Bowen, Antonio Rodriguez, Laura Ray, Chris Kymn, Khari Jarrett
Dartmouth
Richard.Granger@gmail.com

Abstract

The elemental constituent functions of human minds are not yet known, and the paths to identifying these basic "cognitive acts" are constrained at each end by biology and behavior. A coherent architecture derived bottom-up from brain circuits is proffered. We posit principles and pose questions about architectures, their composition, and their applicability to a described range of formidable tasks, with the primary intent of aiding in setting guideposts and challenges for ongoing architecture studies.

Introduction

One researcher may claim that, e.g., categorization is a fundamental psychological operation ("cognitive act"), and hierarchical clustering is a composite. Another may claim (perhaps seemingly counterintuitively) that hierarchical clustering is the primitive, and that non-hierarchical categorization is a special case that falls out in certain conditions. (For our purposes, we will say that the scientific question of identifying the actual primitive elements of human cognition is in fact important.)

How can opposing claims of this kind be tested? It may intuitively seem that categorization must be primal: it is a constituent of hierarchical categorization, so hierarchies must be composites. Yet there are specific brain circuits whose analysis has led to derivation of a hierarchical clustering operation arising directly from the normal physiological operation of those circuits. Without reference to brain mechanisms, the decision of what is elemental and what is composite; i.e., what is the base instruction set, is in principle underspecified (Granger 2006; 2011).

The important thing is to recognize that there is a right answer. The answer isn't known, yet, but that's just "yet". There's a right answer, and it can be investigated.

Past work in our lab and others' has attempted to extract algorithms from individual circuits, e.g., thalamocortical loops, basal ganglia, hippocampal fields CA3 and CA1, as well as from interactions among those circuits, including, e.g., cortico-hippocampal and cortico-striatal loops (Chandrashekar and Granger 2013; Rodriguez and Granger 2016). In each case, detailed algorithmic statements have been derived, software systems have been built, and, wherever possible, side-by-side comparisons with comparable algorithms have been tested.

More importantly for the present paper, these systems have been throughout treated from the perspective of a unified cognitive architecture, rather than as a set of independent free-floating algorithms. The goal of analyzing each individual brain circuit has always been to understand it not just in isolation (which it never is) but rather in concert with the rest of the telencephalic / forebrain circuitry in humans. Each circuit is intended to be studied for what it confers, individually and in integrated loops, to the overall computation of cognition (see, e.g., Granger 2006; 2011). A unified architecture has been posited, incorporating the constituent circuits and their interactions, outlined in Figure 1.

The architecture has been described previously in partial form (Granger 2006; 2011), and a list of educed component algorithms has been posited. These were not arrived at from any first principles nor from considerations of any behaviors. Rather, they arose solely from simulation and analytic treatments of the circuits that comprise the structures in Figure 1. Table 1 lists the resulting proposed instruction set that so far arises from these analyses.

Algorithms derived from telencephalic circuits:
thalamocortical loops ("core"): clustering; hierarchies
" " ("matrix"): sequences; chaining; hashing
striatal complex / basal ganglia: RL/TD learning; bridging
hippocampal fields: time dilation; stacks; match-mismatch
amygdala nuclei: filters / toggles
Shared data structures:
nested sequences of categories (of sequences of categories) = grammars

Table 1: Summary of algorithms and data structures derived from mammalian forebrain circuitry (see Granger 2006).

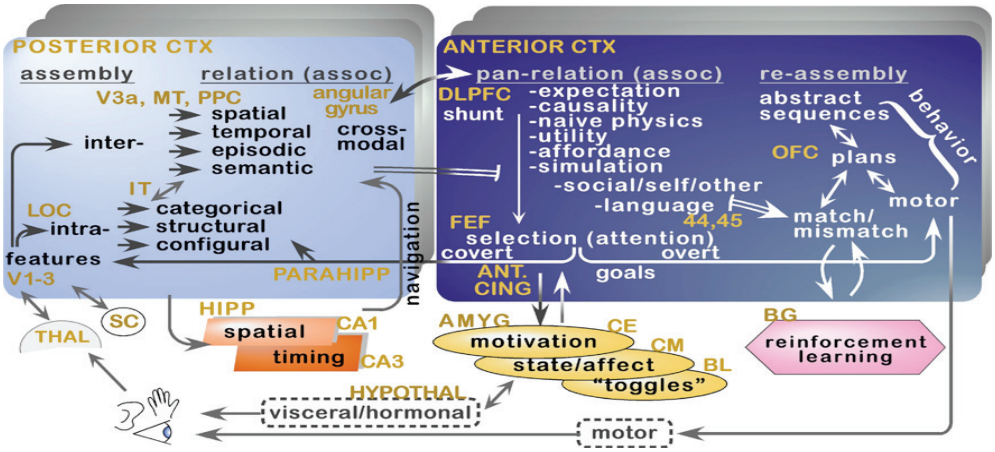

Figure 1. Components of proposed brain circuit architecture, containing all major constituent systems in mammalian forebrain.

Principles of cognitive acts, and examples

We will focus on both principles and examples: principles of cognition and the resulting desiderata (both positive and negative) for cognitive architectures; and examples of tasks that are currently challenging for cognitive architectures, which would be addressable by attending to the articulated principles.

Cognitive tasks are unspecified. We have neither definitions for them nor examples, other than observations of the behavior of humans (and other mammals). Notably, as of 2017, there are no definitions of "human-like" minds, and no instances of them, other than humans.

When a proposed simulacrum is presented, we measure its performance solely by example: by reference to observed human behaviors. Is Siri's language use "human-like"? It can pleasantly surprise us with utterances that could have come from humans (and in many cases did, given the substantial portions of it that are hand-constructed), and it can unpleasantly surprise us with completely non compos mentis responses to queries that a six-year-old would readily have fielded.

The only reason that we know that siri can be outperformed is that humans do so. There is no specification of its intended operation. There is no possible external validation of its performance other than by reference to observed human behavior – not to specifications, which do not exist.

The notion of a cognitive "standard model" refers aspirationally to physics, whose standard model is claimed to be "not a direct model of the entire physical world, focusing as it does only on the relatively low level of particles …." while being "grounded" in levels below and providing a "critical foundation" for levels above (Laird et al., 2017). To truly adopt a physics approach, it is crucial to attend to the key empirical grounding of physics: any theory must be tested for its predictive and explanatory power. If a "theory" doesn't accord with a physical phenomenon at any scale, from quanta to galaxies, the theory is known to be wrong, and physicists focus on eventually fixing it, or supplanting it. For cognition, we may *temporarily* choose a "level" (or time constant or implementation detail) of focus but acknowledge that empirical phenomena at any level are full constraints on any correct theory. The reason that physics focuses on the smallest known components is that all phenomena arise from them. Cognitive theory requires nothing less: the mind is what the brain does.[†]

Cognitive acts are brain circuit operations

Since cognitive tasks have no formal specification other than what humans do, and since the mind is what the brain does, the most pragmatic path to cognitive architectures may be development of algorithms derived from brain circuits.

Deriving cognitive acts from brain circuits is sharply distinguished from saying "we should implement purported cognitive operations in artificial neural networks". Why? Because there are actual anatomical brain circuits, and actual physiological operations thereof, and they are easily shown not to conform to current ANN models (with their (deep, convolutional, or recurrent) variants: DNNs, CNNs, RNNs). (An intriguing minor exception is that of rein-

[†]("The mind is what the brain does" in no way implies that, for instance, "we therefore should not over-treat mental illness just with drugs." Even when we eventually understand how brain operations cause minds, it will still be the case that mental illnesses should not be solely treated with drugs, not because minds do not equal brains, but because mental activity (brain activity) can itself act on the physical brain, and thus can change the brain. Arguments about the powers of thought and reflection are orthogonal to the equivalence of mind and brain. Addressing such category errors is beyond the scope of the present paper, but is important to state clearly.)

forcement and temporal-difference learning (RL/TD), which appear to be very good models of a subset of mechanisms within the basal ganglia, which are called as adjunct subroutines to the primary operations of the cortex).

The assortment of architectural layouts across brain structures, although richly diverse, is nonetheless rigidly constrained by a range of relevant data (Granger '11):

- allometry shows the astonishingly inflexible limits to architectural designs across all mammals, with humans exhibiting no exception (Finlay'95; '05; Herculano'12);
- repetition of cortical-subcortical circuitry designates extensive shared components evidently underlying cognitive operation from low to high levels;
- the disturbingly low precision and slow speed of neural components radically restrict the power of elemental operations; and yet:
- intrinsic parallelism of brain circuits predicts the apparently high Amdahl fractions of the emergent algorithms.

Taken together, and taken seriously, these have given rise to a circumscribed "instruction set" of derived elemental operations from which all complex perceptual and cognitive abilities presumably may be composed. It is assumed that, although the operations so far identified are incomplete and insufficient, the aim of the enterprise is to complete them and test their explanatory power against a full range of cognitive operations. The set so far amassed (Table 1) are all directly derived from the structure and operation of particular circuits and interactions among circuits (Rodriguez et al., '04; Granger '06; '11). Intriguingly, many are not typically thought of as primitive cognitive acts. Ongoing empirical testing is aimed at distinguishing among competing hypotheses, and correcting erroneous ones, as suggested in the introductory comments.

Brain circuits do not resemble ANNs

Current artificial neural network (ANN) "deep learning" and related systems represent a surprisingly modest subset of brainlike algorithms, perhaps accounting for the current wide gap between the capabilities of even the most advanced extant artificial systems versus human capabilities in, e.g., rapid learning from few instances, learning by being taught, attentional mechanisms, navigation, structure, temporal sequences, and semantic language meaning – i.e., in most realms other than statistical "big data" analyses, such as in search-based games.

Any model of a real phenomenon selects characteristics of the real object of study and constructs a simplified version, by omitting or simplifying some characteristics. What's retained and omitted in ANNs? Retained are precisely the following four features of brains: i) very simple individual unit (neuron) computation (add, multiply); ii) messages are scalars; iii) massive parallelism; iv) learning via connection changes. Even within those features, the details are divergent: for instance, error correction does not occur in brain circuits other than in the cerebellum – not in cortex, nor striatum, nor hippocampus – and yet error correction underlies all current major ANN learning rules.

All other features of brains are omitted, and the target question is whether those omitted characteristics either add to or alter the interpretation of the resulting "model." The answer is not controversial. It is easy to show that, for instance, changing the learning rule of an ANN radically alters its behavior and capacities; likewise, it has been shown via models and simulations across myriad labs that different architectural arrangements of neurons, for instance, may entirely change the emergent computation of the resulting system, (as changing the architecture of electrical circuitry unsurprisingly does likewise).

Proposals that backprop-based systems, or systems lacking anatomical architectures occurring in brains, are "models," begs the question of what a "model" is – a model is intended to be a model "of" something. A statistical learning system branded as "cortex-like" or "hippocampally-inspired," without correspondence to the actual network designs of the cited brain circuits, appear to be described in these reminiscent neural terms after the fact, rather than being modeled from the substrates alluded to.

Taking an existing set of proposed "cognitive" operations and implementing them via statistical learning or ANN tools, then, may not aid in identifying primary mental operations from brain circuitry.

Cognition, big and small

Current cognitive architectures are intended to conform to psychological data, specifically limiting themselves to events that unfold over the course of at least 50 (and typically more than 100) milliseconds – moreover, they are aimed at "deliberative" cognitive acts, as contrasted with supposedly lower-level operations.

Some architectures approach this "blurring" of more rapid internal sequences by implementing longer-duration cognitive acts in smaller packages constructed from statistical learning / ANN tools; as mentioned, current ANNs do not contain the vast majority of brain circuit characteristics, and it remains unknown if these omissions fundamentally revamp what actually goes on in brains.

Somehow, real brains give rise to real-world perceptual processing as well as to "deliberative" thought; both low and high level cognitive acts. Might the former derive from highly specialized front end modules, that can be neglected with respect to high level cognition? For the pure initial interface between physics and brain, there indeed exist specialized peripheral structures (retina, cochlea) whose exotic architectures do not resemble those of other brain areas. Once past these very early non-cortical mod-

ules, however, evidence strongly suggests repeated and unified systems that are remarkably widely shared from low-level to high-level regions. It of course could turn out that future research attempts will show these circuits to differ in important ways from each other; at present, such attempts are surprisingly rife with negative results. Moreover, searchers for differential circuitry (or cell types, or genetic alleles) between humans and other mammals yield extraordinarily and unexpectedly disappointing results thus far. The differences found to date appear to fall radically short of any explanatory power for our substantial human cognitive advantages. As these findings continue to stack up, the possibility looms larger that they are not negative findings, but positive ones: our brains are indeed far, far more similar to those of other mammals than we suspected, and our unique abilities (largely language) may well arise from scale rather than differential design (Finlay and Darlington '95; Striedter '05; Herculano '12; Rodriguez '16).

Time, latency, and successive processing

With these formidable architectural constraints in mind, how would one test proposed constructs for their actual human abilities? We forward examples from a broad range of highly replicated experimental paradigms, intended to probe the distinctions between candidate architectures.

Basic levels exhibit fixed latencies

As shown by Rosch and replicated extensively, humans are faster to correctly identify a robin as a bird than as a robin. Moreover, the successive responses (bird → robin) are reliably shown to be separated by seemingly quantal reaction times, adding roughly 100 msec latencies for each subordinate step (Rosch et al., '76). The suggestion has repeatedly been made that these lock-step behaviors reflect a regular timed mechanism.

Cortico-thalamic loops, which are indisputably at play during basic-level behavior, exhibit a timed loop behavior that, in modeling work, gives rise to an unexpected succession of category-subcategory cortical responses, utterly absent from any of the materials from which the models were constructed. The successive basic-subordinate behavior of the model surprised its builders, and suggested that this hierarchical clustering might be a fundamental property arising from the operation of cortico-thalamic circuitry (eg Rodriguez 04). The unexplained reaction time findings of basic levels should be explained by (not just consistent with) a candidate cognitive architecture (Rodriguez '04; Granger '06; '11).

Procedural-declarative dissociation

Studies have investigated the potential application of cognitive acts in architectures (e.g., ACT-R) in problem-solving tasks such as tower of hanoi. An additional question of interest is the finding that selective damage to the medial temporal system enables the tasks to be learned, yet dissociates the learning from any retrievable memory of the task (tower of hanoi, mirror writing, and many other "procedural" tasks conform with these findings). Architectures have been constructed with "declarative" and "procedural" component modules, conforming to the differences. It may be of interest to identify explanatory accounts of how these modules differ from each other, and predictive accounts of how they arise from starting principles.

Implicit associations, fast and slow

When asked to pair images with descriptive adjectives, subjects exhibit slower reaction times when the pairings are at odds with their internal biases. Thus, pairing african-american faces with laudatory adjectives is slower than pairing these images with negative adjectives – this occurs both for caucasian and african-american subjects alike. When debriefed, subjects typically claim no such bias, and in fact may profess liberal leanings, while nonetheless exhibiting the differential RT behavior described. Indeed, once told about the test, subjects re-taking it then exhibit slower responses to all stimuli, whether or not the pairs reflect biases. These speed reductions are accompanied by increased activity in anterior cortical areas, presumed to be activated later than initial regions, subsequently inhibiting the initial (biased) response. An explanatory account may presumably indicate both the early (biased) and late (intentionally corrected) responses. Much of the panoply of examples described as "fast" and "slow" (Kahnemann '11) exhibit related characteristics.

Perception is not solely feed-forward

Many early-stage perceptual phenomena are processed more rapidly than 100 msec; these may perhaps thus be judged to fall outside the proper purview of current cognitive architectures. Researchers often add specialized "front end" or modality-specific modules to a given architecture to produce "hybrids" – part specialized perception module, part "deliberative." As discussed above, neuroscience studies find that, other than the extreme periphery, the processing of sight and sound is carried out by circuitry very similar to, and highly integrated with, the rest of the brain, both for feedforward and feedback communication. One might reflect on how much processing underlies the integration of, say, a dog's appearance and barking sound. Are there specialized modules not just for vision and sound, but also for their cross-modal integration? Or is the early processing of images and sounds just another in a long line of cortical regions seamlessly and successively operating on ever higher representations of inputs? This is just the kind

of question that some in neuroscience attempt to study, and that cognitive architecture research could contribute to.

A side effect of assuming separate modules for vision is that some modes of perceptual processing become much more or much less easy and natural as a function of the assumption of the division of labor in the architecture. If there is a separate "vision" front end, followed by "deliberative" reasoning about the visual input, we may think that the "vision" component has "completed" its job, providing input to downstream modules.

Figure 2. Successive images extracted from a video sequence, illustrating how much features change rapidly over frames.

These views of a truck (Fig.2), separated by just a few frames, have very few shared feature characteristics. An independent "front end" that attempted to recognize them would be highly unlikely to rate them as similar, let alone to categorize them as the same object. Yet due simply to their contiguity over time, the visible portions of the truck can be recognized (by us) as being the same object, and even the relationships among its constituent parts can be registered. The confusing variability of the bottom-up / feed-forward features can be overcome by top-down / feedback information, when an architecture's perceptual and memory processes interact. Front end systems (such as SIFT) are not only computationally expensive, they have repeatedly been shown to give incorrect labels since they do not have feedback from memory, motion, contiguity, etc., whereas the low-level input features (such as the truck) are utterly different once the object has moved, or turned, or is in different lighting. Following the most prevalent visual processing methods designed for still images, the task of unifying the view of the object would be made artificially (and unnecessarily) difficult.

We argue that these are not simply specialized issues concerning visual processing; they are issues that are integral to cognitive architecture design: the need for feedforward and feedback information; the need for integration across successive processing stages; the need for perception and memory to interact. Many candidate architectures contain characteristics of this kind, and yet many would nonetheless find Figure 2 challenging, suggesting that it may be fruitful to search for further integrative processing that could appropriately treat such natural everyday tasks.

Detailed explanation of our approach is perhaps too specialized for this paper, but we provide just a brief account in hopes of illustrating the key architectural points that we are advocating. The architecture uses models of the two primary retino-thalamic pathways from eye to brain: the parvocellular and magnocellular paths, which can be loosely thought of as, respectively, static image based, emphasizing contrast (parvo), and time based, emphasizing motion and contiguity at the expense of detailed static feature identification (magno). In particular, the magnocellular pathway predominantly tracks motion that is relatively consistent over time, without searching for wholly consistent features (as, e.g., SIFT would be forced to do).

The truck in these images is identified and tracked over time largely via contiguity. No expensive front end such as SIFT, nor tracking via optic flow (also expensive and typically proceeding via very few frames at a time), are used. We posit that low-level input features are inherently neither predictable nor reliable, which (if they were used) would lead to inherently questionable inferences. The result in our case is an extremely inexpensive front end that is designed, architecturally, to be used not as an isolated module but in concert with downstream mechanisms designed to evaluate successively longer time spans, and thus no longer dependent on still images, nor front end modules that must recognize in isolation (Bowen et al., 2017b).

When do two things look alike?

As traditionally "front end" modules are recognized to actually be integral parts of the architecture, additional tasks can be seen included in the evaluation of the approach. An example is the deceptively simple question of when two things look alike. There is a substantial literature, entailing a range of issues including analogy, contextual setting, and many more topics, but we have recently asked the question at a surprisingly simple level, with encouraging results.

The general form of the question admits inputs that could range from visual or auditory percepts up through complex abstractions. Taking just the simplest example, if we consider two images, the task of "image quality assessment" (IQA) is one that is wholly dependent on human behavior: the only measure of "how similar" two images are is defined as "whatever humans do".

It is notable that this is a case in which there is no external specification other than human performance, and yet human performance can be measured with precision, so candidate approaches can be impartially evaluated. (Would that there were more such instances!)

We give a very brief precis of work by Bowen et al (2017a). Subjects are shown pairs of images and asked to rate their similarity from 0% - 100% similar. Image pairs can be generated by producing degraded versions of an image via JPEG compression. The "right" answer is solely "whatever humans do" – measures are evaluated in terms of "difference mean opinion scores" (DMOS), i.e., the average rating given by humans tested on particular images.

The field of image quality assessment strives to predict what these human scores will be, from the images alone.

Traditional approaches simply measured root mean squared distances, pixel by pixel across the two images. That approach typically achieved passable estimates, with plenty of room for improvement. The current state of the art is a family of methods termed "SSIM," based on a multiplicative average of three measures of multi-pixel comparison across two images (Wang et al., '03). SSIM reliably outperforms Euclidean root mean squared distances.

We have recently shown that a method derived directly from brain circuit models, as part of the architecture depicted here (Fig 1), outperforms the current industry standard (SSIM). The brain circuit derived method predicts human evaluations of image similarity better than the current state of the art method. Moreover, the method is also applicable to abstract similarity evaluations used by multiple researchers (Attneave '50; Medin '78; Nosofsky '92). This is simply a suggestive reminder that incorporating brain considerations, and inclusion of characteristics from low level to high level, may aid in arriving at cognitive architectures that actually match human performance.

Whither cognitive architectures?

Architecture design, then, may benefit from the inclusion of cognitive acts from front to back and from fast to slow, including those that traditionally have been excluded from "cognitive" architectures. The inclusion of the full range of cognition may highlight structural issues that serve as useful constraints on architecture design. Constraints, whether from neuroscience, from reaction times, from comparative anatomy, from allometry, from computational costs, all provide indispensably useful guidelines helping narrow down the space of eventual architectures that could not just match human behavior, but provide predictive and explanatory accounts of how it arises.

Human, or "human-like"? At present we have no examples of "human-like" minds whatsoever. Perhaps worse: we have no specifications(!) and thus no means for measuring how "like" a mind is to ours, nor for falsifying a hypothesis that a candidate mind is or is not "human-like". Are there "human-like" minds? How would we know?

Deliberative, or reportable? Elemental cognitive acts are all (or largely) unreportable. Should this make them ineligible as the bases for cognitive architectures? The elemental cognitive acts proposed here:

- seem not to fit a definition (which does not exist) of the intuitive term "deliberative";
- mostly are more rapid than 100 msec;
- are not subject to intuition nor overt inference;

Seeking to avoid the problem by avoiding the low level may be steering the enterprise in a misleading way. We have suggested that inference, bias, even widely used (but quite poorly defined) terms such as recognition and categorization, all entail mechanisms that are unreportable; ought they all be omitted from cognitive architectures?

Modules or successive stages? The present approach argues that, based on brain circuit design, it is the existence of successive processing stages communicating with each other, that enables seamless interaction from percept to concept. Is this a desideratum for cognitive architectures?

References

Attneave F (1950) Dimensions of similarity. *Amer Journal of Psychol* 63:516-556.

Bowen E, Rodriguez A, Granger R (2017a) The perceptual curvature of human image quality assessment. (In prep).

Bowen E, Jarrett K, Ray L, Kymn C, Granger R (2017b) Space and time costs in a hierarchical architecture for video. (In prep).

Finlay B, Darlington R (1995) Linked regularities in the development evolution of mammalian brains. *Science* 268:1578-1584.

Chandrashekar A, Granger R (2012) Derivation of novel efficient supervised learning algorithm from cortical-subcortical loops. *Front. Comput. Neurosci.,* 5: 50. doi: 10.3389/fncom.2011.00050

Granger R. (2006) Engines of the brain: The computational instruction set of human cognition. *AI Magazine* 27: 15-32.

Granger R (2011) How brains are built: Principles of computational neuroscience. *Cerebrum; The Dana Foundation*. http://dana.org/news/cerebrum/detail.aspx?id=30356

Herculano-Houzel S (2011) The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proc Nat'l Acad Sci* 109:10661-10668.

Kahneman D (2011) Thinking, fast & slow. *Farrar Straus Giroux*

Laird J, Lebiere C, Rosenbloom P (2017) A standard model of the mind. *AI Magazine* (In press).

Medin D, Schaffer M (1978) Context theory of classification learning. *Psychol. Review* 85:207-238.

Nosofsky R (1992) Similarity scaling and cognitive process models. *Ann Rev Psychol* 43:25-53.

Rodriguez A, Whitson J, Granger R (2004) Derivation and analysis of basic computational operations of thalamocortical circuits. *J. Cognitive Neurosci,* 16: 856-877.

Rodriguez A, Granger R (2016) The grammar of mammalian brain capacity. *Theoretical Computer Science C (TCS-C)* 633:100-111. doi: 10.1016/j.tcs.2016.03.021

Rosch E, Mervis C, Gray W, Johnson D, Boyes-Braem P (1976) Basic objects in natural categories. *Cognitive Psychol* 8:382-439.

Shepard R (1962) The analysis of proximities: Multidimensional scaling with an unknown distance function (II). *Psychometrika* 27: 219-246.

Striedter G (2005) Principles of brain evolution. *Sinauer Assoc.*

Wang Z, Simoncelli E, Bovik A (2003) Multi-scale structural similarity for image quality assessment. *Proc 37th IEEE Asilomar Conf on Signals, Systems, Computers.*