

**The Hybrid Vigor Journal
V1.3:04-2002**

**© 2002 The Hybrid Vigor Institute
To be freely copied and distributed
only in its entirety, with all credits intact**

**AS IF YOU WERE THERE
Matching Machine Vision to Human Vision**

**Richard Jay Solomon
University of Pennsylvania
Program on Vision Science & Advanced Networking**

**The Hybrid Vigor Institute
<http://hybridvigor.org>
<http://hybridvigor.net>
Human Perception**

April 2002

CONTENTS

[0.0] Abstract	3
[0.1] Acknowledgments	4
[1.0] Introduction	5
[2.0] Key Human Visual Parameters	8
[2.1] A more utilitarian model of the visual system	11
[3.0] Old Models of Vision.....	13
[3.1] Defects of conventional TV	15
[4.0] A Better Model for Electronic Motion Imaging	17
[5.0] Some Surprising Research Results About Vision	19
[6.0] Conclusion.....	20
[6.1] Proposed Research	20
[7.0] About the Author: Richard Jay Solomon	22
[8.0] Endnotes, Bibliography and Further Readings	23

[0.0] Abstract

A large variety of advanced electronic imaging equipment is available for collecting and disseminating visual information. However, despite the ever-expanding capabilities of new devices, the limiting factor for understanding and reacting to information displayed and collected is the human perceptual system.

Where science has had any discernable input (other than guesswork and trial-and-error) in the design of photographic and television systems, parameters for imaging systems have been set primarily via psychophysical measurements. Psychophysics, the study of human reactions to physical stimuli or input, is limited to determining reactions from external stimuli; it does not study how the brain works. It is difficult to explain just how the human perceptual system reacts to these stimuli. What has emerged as a key design problem for a system that more accurately replicates “presence” is that much of the psychophysical data used in the past to engineer high-performance networked imaging systems — i.e., simple reactions to stimuli — is not consistent with the known workings of the human neurological system, which tries to explain *how* or *why* we react.¹

The latest neurological data, using direct brain scanning devices such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) scans², indicate that the human sensory system is much more sensitive, yet at the same time is much more selective to information stimuli from visual displays and auditory sources, than had been previously understood by engineering designers. Seemingly contradictory, these dual insights imply that the designs of information transmitting, storage and processing devices have to be more closely coupled to the human perceptual system in order to gain a better “impedance” match between opto-electronics and our neurons.³

Compared to older theories based on psychophysical measurements, many of the results recently published in the neurological literature about how the human vision system works are surprising and counter-intuitive. This new research forces us to question long-held assumptions about how electronic transmission components, cameras, displays, processors, and even audio speakers should work. We can use this new information to design much more accurate and believable electronic systems that would replicate a scene as if the observer were present.

That is to say, for critical scientific, medical, archival and engineering objectives — contrary to the design of consumer entertainment appliances — it is simply not acceptable to discard potential perceptual inputs to the human neurological system based on erroneous ideas of what we can perceive and not perceive, even if the picture looks “pretty good” to the untrained observer, or at least good enough based on what we’ve become accustomed to,. The human vision system is much too complicated and capable for us to settle for simplistic design objectives such as those found in most off-the-shelf compression, transmission and display systems. Making pretty pictures for television and movies is easy compared to providing critical information for technical analysis and archival storage.

[0.1] Acknowledgments

I wish to thank my colleagues Eric Rosenthal of Creative Technology LLC, David Farber at the University of Pennsylvania, and Tice de Young of the National Aeronautics and Space Administration who kindly contributed some of the concepts in this paper. I take responsibility for all misinterpretations, however.

[1.0] Introduction

No one has ever mistaken ordinary television or motion picture systems for reality. No one, that is, until my colleagues first turned on an experimental high-definition television (HDTV) camera that we were working on at MIT a few years ago and fooled me into thinking that I was looking into a glass diorama, in this case at a vase full of flowers. At first I thought they were playing a game with me, back-projecting a 35mm comparison slide on a screen — that is, until someone reached in front of the camera and moved the flowers. Startled, I rushed to the screen, and asked, “Where are our scan lines?”⁴ That’s what the engineers at Polaroid, who with Philips and IBM built the new camera system to our specifications, also wanted to know: Where did the scan lines go? Why does this look so real and not like television at all? We expected our designs to look better than any other HDTV system to date, but not so good we’d fool ourselves!

We had an research project funded by NASA (the U.S. National Aeronautics and Space Administration) and DARPA (the U.S. Defense Advanced Research Projects Agency) to demonstrate an electronic full-motion imaging system⁵ that could be used to capture remote images without the artifacts, or imperfections, commonly found in off-the-shelf consumer HDTV, such as small blocks in the picture (called “pixilation”), or the flickering interlaced scan lines, or the distorted color. We had surmised that conventional television designs⁶ would not suffice for an image analyst, a scientist, a medical practitioner, or anyone who needs to see something at least as well as if they were present at the scene.

HDTV, in the format being hyped by the consumer electronics industry, eliminated too much information (and added a few extra things as well) in order to meet low-cost marketing criteria, yet make the pictures prettier enough to sell at a premium over the cost of standard television. Anything that removes information from an image, even subliminal stimuli that consumer systems tend to downplay — such as scan lines, “grain” on film, or the jump and weave of motion pictures which you do not usually see unless you pay attention to it — reduced its utility to our demanding sponsors. Hence, most compression-driven HDTV systems, though yielding perfectly adequate pictures for entertainment, are precisely what our DOD and NASA sponsors were not interested in.

The guiding principle for consumer HDTV has been, “how little information do you need to transmit to make the picture look good,” not “how much information can we capture and transmit

to give the viewer what he or she would see if present at the scene” (or, for that matter, *more* than the naked eye would see if present). The reasoning behind that principle is quite simple: transmission channels⁷ are expensive (for the consumer market) and the U.S. government's mandated goal was that HDTV had to re-use the existing over-the-air spectrum without requiring more channel space than that allocated for conventional TV. As it turned out, even re-allocating the existing TV spectrum to HDTV has become a political football, since broadcasters who control the spectrum do not want to allow anyone else to use it, even if they aren't (recalling Stalin's view of the world, "What's mine is mine, what's yours is negotiable"). But various visual tricks have been employed to make the picture better, most of them by digitizing the analog signal and cleaning up some of the "noise"⁸ inherent in analog transmission.

Using our intuition, and a few critical studies from various disciplines that indicated certain subliminal cues should not be eliminated, we had a camera built by Polaroid, Philips and IBM as subcontractors that transmitted live, color, non-compressed, *analog* signals to a standard graphics-art (computer) monitor, not a TV set, at approximately twice the spatial resolution of a typical American television set, and with double the "frame rate" — meaning twice as many images could be sent in the same period of time, displaying motion more realistically as well.⁹ Most important was that the frames were not interlaced, as in conventional television, and were captured coherently — that is, all the pixels were captured simultaneously like film, instead of being scanned over a 1/30 or even 1/60 of a second; it turns out that the human visual system is sensitive to such scanning delays, even though most television engineers may think otherwise.

This was the first time that humans had ever seen an electronic image of such high quality, and the result was the unexpected phenomena of observers thinking they were looking at a real object, not a picture. We had similar experiences with almost everyone who saw that demo the first few weeks. The most interesting when we brought the camera and display down to the National Institutes of Health, and several scientists reflexively tried to touch the objects on the screen. One of these, a neurologist, asked me what we were doing to create such realism, and my rejoinder was, "Tell me how the human vision system works, because I don't know why this looks so good myself!" He handed me a reading list, and said he hadn't a clue.

That inspired a multi-year project to attempt to discern the factors necessary for an imaging system that would more realistically represent actual "presence". After traversing several disciplines, including neurology, neurophysiology, psychophysics, quantum mechanics, physics, biochemistry, biology, electronics, ophthalmology, applied optics and mathematics, photonics

and acoustics, we have more of clue today. But we still have much more to learn about human vision in order to develop a family of devices that may enhance visual reality for surveillance, scientific investigation, engineering design, remote medical procedures, and archiving rare art works, among other critical applications.

[2.0] Key Human Visual Parameters

Several factors have been identified by neurologists, such as Prof. Semir Zeki, of the Institute of Cognitive Neuroscience at University College London, as key to an understanding of how we see and how we use visual information in decision processes. (Zeki was among the first to apply perceptual stimuli and observe the brain's reaction using modern tools, such as functional MRIs, to begin to understand how the vision process works — in particular, how the brain makes the world appear constant around us.) The most important factor is our vision system's ability to maintain color constancy despite ambient light shifts. When we look at our skin in incandescent room light, and then in the light coming from the window, we don't see a radical shift in color even though physical measurements (or the output of film or a video camera) would detect a strong reddish tint from the room light and a strong bluish tint from the skylight. Even more interesting, when you put on yellow sunglasses, for a few moments everything is yellowish, but almost immediately you see your environment as if you were not wearing tinted glasses at all. Testing with accurate color measuring instruments shows that the tinted glasses have almost no effect on our perception of accurate color. Why? Well, we really don't know, but clearly some powerful processing is taking place in the brain's cortex that is not replicated in any device yet invented.

This might seem even odder when we consider another critical human visual factor: our extreme sensitivity to spectral (i.e., color) differences. As Patricia Churchland, a philosophy professor at the University of California, San Diego, first observed, spectral sensitivity can be thought of as a parallel cortical process to our ability to resolve tiny spatial differences; that is, our sensitivity to where objects are located in our physical environment. We can see things with our lens and retinal systems that conventional optical theory says is impossible, such as our demonstrated ability to see a fly land on a telegraph wire at a very far distance in the desert. This ability, termed *hyper-acutance* (acutance being perceived "sharpness"), is a well-recognized visual phenomena by ophthalmologists. Churchland, who pursues her interest in perception in parallel to her studies of philosophy, demonstrated that the same mathematical simulation of the phenomenon that can accurately predict human *hyper-spatial-acutance* also predicts *hyper-spectral-acutance*. (Spectral acutance has generally been ignored by psychophysicists as a phenomenon.) This may mean that the same brain activity is responsible for both types of acutance.¹⁰

Motion also is critical to human visual perception. Since our eyeballs, hence our retinas, continually move, there is no such thing as a totally still picture. Indeed, as your eye doctor will tell you, if you cause the saccadic motion of your eyeball — the rapid, intermittent eye movement occurring when the eyes fix on one point after another in the visual field — to stop by applying pressure on an adjacent nerve or through drugs, you will rapidly go blind. This is a reversible phenomenon, though I don't recommend trying this by yourself without medical supervision.

Saccadic motion is not random, but is programmed by the cortex, an indication that the shaking eyeball is inherent to the process of vision. (There is a physio-chemical reason for this shaking, but just how the retinal structure is sensitized to photons is beyond this paper, and is not necessary in order to understand how our vision system works.) Furthermore, and most important, the saccadic motions are not linear, but follow complex, “pre-programmed” macro and micro motions. So even if an image can be tracked to follow your retina by some external device, the micro-motions jiggle the cones and rods, and no image can be perfectly static. This turns out to be a very useful scanning mechanism for vastly increasing our sensitivity to subtle visual phenomena, and is the basis for hyper-acutance, but also sometimes for confusing us with subtle illusions, of which the textbooks show hundreds.¹¹

These fundamental, and somewhat paradoxical biological phenomena — extreme sensitivity to color yet with strong color constancy under varying illuminations, and continual visual motion with hyper-acutance for static objects — work in conjunction, and in ways that are not accounted for in conventional television and motion picture systems. Color influences the perception of moving objects, and motion influences color. Most image compression technologies, which are widely deployed to squeeze a visual image sufficiently to be transmitted over slow transmission channels, simply ignore these phenomena, leaving our brain to guess what is missing.

For example, a well-known scientist (who has not yet published his results, thus the anonymity) was surprised to recently learn, using non-invasive brain scans as his subjects were observing real scenes, that if color cannot be recognized within a few frames, the object itself will be wiped from visual memory. Since most compression algorithms modify certain colors that had been believed to be “unimportant” according to standard psychophysics, such compressed electronic television systems are anathema to an image analyst attempting to ascertain important details. This could be a cancer cell, or a deliberately camouflaged tank, or perhaps the subtle motion of

an object on a distant planet. Now it's there, and with incorrect compression, *zip*, it's gone. Deleting motion and spectral changes, although the image appears to be satisfactory, may delete critical, *subliminal* cues that might lead to important perceptual information. In science it is better to keep everything and not delete until final analysis — and maybe not even then, if retrospective work is anticipated.

Even steady-state spatial scenes can be enhanced by motion. We have demonstrated that by merely taking a very high-resolution image and repeating it with ever-so-slight movements — or better, a short sequence of “stills” — it doesn't just look sharper, *the repetition actually adds more information* to the visual precept. Discarding psychophysics, and thinking about this neurologically, what is happening is that the cortex is integrating spatial information over time. Some yet to be understood neurological process(es) make us think we see the entire very short sequence simultaneously, with each temporal scan adding bits of information to the whole.

The result of this integration is that the image we really see greatly increases in acutance (or “sharpness”) over what a theoretical, simple static image would give us. Standard optical physics cannot account for this, as we noted — indeed, it has been well-known for a long time that we can see details in the distance about 50 times greater than what optical theory predicts for our combined lens and retinal structure. A puzzlement, until one considers the complexity of an eyeball that shakes (saccadically) according to a cortical feedback algorithm, and a brain that picks out the details and knows what to overlap and what to discard. This is a *real* neural network, and it doesn't follow anything we yet know about computation.

When we realized that the human visual system is not comparable to the way we have been designing cameras — what an electronic engineer would call a large *impedance mismatch* between the human vision system and the electronic systems — we understood we had to return to basics. At that point, I began to study how and *why* the human vision system perceives images, just to understand why our special camera worked as well as it did. Color, or more precisely, spectral separations, was my first clue, since experiments we had done at the MIT Media Lab some years before our camera was designed indicated that color and contrast were more important for acutance than the number of lines of resolution produced by the imaging devices.

More important, we needed to know the influence of enhanced multi-spectral sensitivities on motion perception, since what we were doing at the time was designing *motion* imaging

systems, not still cameras. (There is no point in even thinking about static imaging, since we've now learned that all perception is in motion.) And we should realize that virtually all systems we use for motion pictures are misnamed: the pictures don't move, we just think they do. And that is the crux of the problem. *We haven't a clue as to why we think moving pictures (film or television) move.* The oft-cited "persistence of vision" theory, assuming any image actually "persists" on the retina, is a description of what happens, not an explanation. If the image were really to "persist", we'd see a blur every time we move (which is all the time). If we account further for saccadic processing, we realize that nothing could persist, yet we integrate sequential images and think either the objects are moving, or as I just described, the objects get sharper — or to use the proper terminology, have greater acutance.

Why is this important? Because the various and complex motions of photons emanating from the display frames of film or television and hitting our retinas are not necessarily in sync with anything going on in our brain. Sometimes, in the case of specific frame rates, bad things can happen — as when a few years ago, an animated cartoon in Japan flashed objects at a dangerous rate that caused epileptic seizures in children. This is called *nystagamus*, and was first observed in the 1840s when people became ill watching telegraph poles flash by railroad train windows at the previously unheard of speeds of 30 mph. People had never moved that fast before. Travelers learned not to stare at poles, and today, the super high-speed trains in Europe and Japan take in account pole spacings and window apertures to prevent nystagamus.

But this is just one small example of a bad impedance match — there are dozens of identified, temporal cycles, or phases, in our cortical loops and its retinal extensions (our eyes are our only sensory devices which are a direct extension of our brain) that work in contradiction to man-made frame rates, video scanning algorithms, compression schemes and other opto-electronic mechanisms that end up degrading our perception of a remote image. Hence, our analogy is to an impedance mismatch; here impedance is defined as an incompatible phase relationship between an electronic system and a human system — something which is inherent in *motion* imaging.

[2.1] A more utilitarian model of the visual system

These insights led us to a process of delineating a new, more utilitarian model of the human visual system. Because of the phases in which our retinas and brains operate, the fundamental baseline of this model must be *temporal*, or motion-related, rather than static or spatial as in most other imaging models. The data for this new model will be derived from current

neurological research which helps explain how the transfer of information to the human perceptual systems may be optimized; the psychophysics upon which current machine vision systems are based, on the other hand, merely measures the information transfer. We need both disciplines in order to improve mechanisms that deliver information to our visual system. Properly designed devices should also reduce information overload by helping humans filter extraneous information, and making significant improvements in visual signal-to-noise ratio.

We can envision such systems being embodied in “intelligent”, electronically-enhanced, adaptive cameras, binoculars, telescopes or microscopes, for example. Not only would one see things magnified, but depending on elements in the scene, the spectrum would be expanded to include the infrared and ultraviolet, or alternatively, to separate the visible light colors beyond what a normal human is capable of doing. Subtle changes in the spectrum could be emphasized or augmented to distinguish surface features not readily noticeable. Spectral, or color, augmentation could be used to enhance acutance or sharpness, as the naked eye apparently does on its own, but which most imaging systems now suppress in order to limit the amount of information that is transmitted. The possibilities for improved vision are manifold once we employ mechanisms that capture and display the full sensory environment instead of only the part that looks attractive. Processing of information should take place at the end of the chain — at the display device — not at the beginning, as it is today in order to conserve transmission bandwidth.¹² For scientific, medical and similar applications, bandwidth constraints are vastly different than for entertainment networks; indeed, bandwidth has become so cheap, we must wonder why so much effort is placed into making it artificially scarce.¹³

[3.0] Old Models of Vision

Conventional imaging systems are based on 150 years of photographic experience. Up to very recently, scientific data used to design and optimize conventional photographic and video systems have been mostly collected via trial-and-error testing of human observers using psychophysical measurement techniques that did not use modern experimental techniques such as double-blind trials¹⁴ and sufficiently accurate sampling methods. Within the last decade, new ways of measuring human perception based on direct feedback from brain processes have been developed using functional MRIs et al.; perceptual data from these new techniques have indicated that many of the critical assumptions about human visual and other perceptual processes based on psychophysical measurements and models require modification. Since many of the components of our photographic and electronic imaging systems, from capture through storage and compression to display and analytic processing, have been based on these older models, it is clear that much improvement can be made in system design by re-evaluating these assumptions.

Virtually all photographic and electronic camera systems are based on a simplistic model of how the human visual system historically has been thought to work. In the older models, the human physiology was presumed to consist of these elements:

1. A lens which focuses an external image on the retina in the eye, and a package of muscles and neurons that controls what the eye looks at;
2. The retina or retinal structure, acting somewhat like film or a photosensitive electronic element, which:
 - a. Counts the received photons, determining luminance (brightness) levels;
 - b. Segments the luminance spectra into three, broad frequency bands (usually red, green and blue), specifying color (chroma);
 - c. Dissects this image spatially, similar to pixels in an electronic camera or silver halide grains in film, and somehow stores this dissected image over time (akin to “frames per (unit time)” in a movie or television camera);
 - d. A vaguely specified mechanism, somewhere inside the human brain, which reconstructs the received external energy, producing a miniature version of a picture to be beheld by the viewer (at one time this was stated as a “homunculus” or “little man”, who recursively viewed the image inside one’s cranium).

The underlying assumption of this model is that for motion imaging it is sufficient to merely build an electronic (or photochemical, or mechanical) version of this image dissector to sample four basic functions: 1) spatial objects or their edges; 2) the level of brightness or luminance; 3) the color reflected or transmitted by these objects; and, 4) the changes over time of the first two functions. Then, all that is necessary to fully represent the data in the distant scene, for a human to see on a display, is to perform linear transforms (that is, the linear correlation between, say, an electrical signal and degree of brightness — if the signal increases, so does the brightness) of the input and output mechanisms.

The implied trick is that if the sampling rate (both in time and in space) and the transfer functions follow mathematical information theory, then it should be feasible to replicate a scene as if one were present. This is the same general idea as when you transform music from analog to digital form: if you sample at a higher rate, you are supposed to get better sound quality. But it's been shown that information theory alone doesn't work for this kind of perceptual information, as we explain below; no matter how high the sampling rate, it never sounds as though you are *there*. (For music, much more needs to be done to replicate "presence", and the same is true for images.)

In reality, it has proven extremely difficult to build any device that actually represents what a human would see *as if that observer were physically present*. Yet, since in many critical scientific or medical environments, human presence is often impossible or awkward (as in small cavities), electronic "presence" would be most desirable. Further, accurate storage of such information would be useful to analyze events retrospectively, to assess inputs from other observers and experts, and to apply advanced information processing tools such as data mining.

There are other critical applications beyond science and medicine where enhanced electronic presence would be extremely productive, or essential: archaeology, where bringing experts to the scene may not be possible; preservation or authentication of artworks, where the original may be lost, or in question; surveillance or inspection of dangerous devices or environments, where a human observer may be at risk; and in courtrooms, classrooms, boardrooms and theaters, where accurate illustration must take place before large groups and first-hand inspection is expensive or not practical.

[3.1] Defects of conventional TV

There are several reasons that conventional television systems (including most HDTV applications) fail to achieve reality imaging or presence:

1. The psychophysical data used in the conventional electronic model are rough approximations based on very small and highly skewed samples of human subjects. Humans have a much wider range of differences in their perceptual systems and systems need to be tailored to these differences where critical decisions are to be made. There is even evidence suggested by genetic research by John Mollon, a professor of visual neuroscience at Cambridge University, that human females have a wider color range than human males; that is women see some colors men do not see, because the genes which define that sensitivity are found only on the X chromosome, and women have two of them while men only have one. As noted, some of the more important assumptions about "average" standard observers contradict recent neurological and physiological human vision system research.

2. Data collected by conventional electronic sensors may be ambiguous — i.e., provide the wrong information, or information too open to subjective interpretation — taken in isolation. Information capturing hue fluorescence, intensity and other surface characteristics are almost completely ignored or discarded, reducing the ability to replicate reality. Humans have a much better processing system to account for ambiguity. It would be worthwhile to emulate nature in this respect, and to augment it with electronics where possible, instead of handicapping the system.

3. Further distortions are added by compromises intended to conserve the limited bandwidth engineered into the television transmission system. Common digital media tools that enhance the edges of objects actually lower information content while giving the illusion of increasing sharpness — an unnecessary artifact with a strong bearing on critical imaging applications where subtle details can be of highest important for analysis, or present the least impediment to an experience of "presence." (Edge enhancement also makes conventional camouflage techniques almost useless. These tend to highlight a mélange of patterns with sharply defined chromatic themes, albeit of a constrained, usually greenish-yellowish, hue — this is perfect for

off-the-shelf digital video cameras to pick up. Bluish colors following a gradual gradation would hide better from conventional cameras.)

4. Conventional television, and most film systems, limit their color space to a subset of what can be sensed and displayed using red, green and blue “lights” (RGB). It has been known for a long time that the human retina can sense a color space much greater than can be captured and replicated with only three broad spectral separations, and the theory that only RGB is necessary to cover the entire visible spectrum can be shown to be both physically and mathematically flawed. In graphic arts, a minimum of 5 or 6 colors is almost always used, and more for extremely high quality reproductions. Furthermore, the human visual system can handle a contrast ratio approaching 1,000,000:1, but the best display systems are limited to less than 1000:1. As a result even the outputs of advanced cameras cannot today be perceived by humans, further restricting visual information analysis, because the display systems aren’t as sophisticated as the cameras.

5. And most important, the incorrect parameters set under conventional psychophysical measurements (such as linear or logarithmic relationships), in terms of actual human perceptual response, have placed artificial limits on media and devices currently used to capture and replicate original scenes.

Recent examinations of the retinal and cortical neuron wiring offer a much more complex model of how we sense color and luminance. This model, still being developed by neurologists, is based on feedback processes between various parts of the brain and the wiring in the retina itself. As noted, neural mechanisms related to color sense have a direct bearing on the detection and recognition of moving objects. This is contravened by most motion imaging systems today which use compression algorithms that suppress certain spectral frequencies. Suppressing these frequencies prevents full exploitation of powerful human cortical motion processing, thereby distorting the available visual information while still making pleasing pictures.

[4.0] A Better Model for Electronic Motion Imaging

Efforts to replicate reality using the conventional (linear) model have been fraught with difficulty. Evidence has accumulated that this is because the generally accepted model is flawed — it is *not* the way humans see or perceive. Our perceptual system is much more sophisticated, more sensitive to useful information, and much more complicated than the camera model implies.

A better model would recognize that audio and other neurological inputs — including memory, torso and head position, and mood, among others — affect visual perception. In our early research, conducted at the MIT Media Lab in the 1980s, we surmised that audio enhanced vision, and ran a set of controlled experiments whereby we displayed not very good television images of a sporting event accompanied by several levels of sound, from just plain awful to extremely high-fidelity audio of the crowds cheering and the general ambient noise of the stadium. We found that the television images appeared to our control subjects to improve as the sound got better, even though the television pictures actually remained the same. This research led to the deployment of surround-sound video products.

Since then, research by John Watkinson, an independent consultant in digital video, audio and data technology based in the U.K., has shown that being able to orient precisely where a sound comes from is an important input to human perception of both sound and images; the explanation from neuroscience is that the processing center for vision also mediates the aural inputs, and as Watkinson has demonstrated with specially designed active speakers, the human cortex is extremely time-sensitive as to what sound arrives in which ear. Indeed, the cortical process for sound orientation appears to be parallel to that for visual orientation.

What we are learning from linking conventional psychophysics experiments to neurological research is that the human visual process is much more subtle than had been supposed. It is not merely a mechanism which simply plots linear transforms from outside energy functions such as light or heat, or detects ambiguous "just noticeable differences," or JNDs, in energy sensitivity (our ability to distinguish between intensity or frequency of energy, such as green v. red, bright v. dark, or sensing infrared or the presence of another object), common errors assumed in psychophysical measurements. There are more critical factors in human vision than merely sampling for luminance (the intensity of light per unit area of its source), and discriminating colors broadly by hue and saturation (the amount of white in a color). Indeed, it

turns out, as the famous psychophysicist J. J. Stevens showed in groundbreaking experiments a half century ago, even those transforms are neither linear like most television systems, nor logarithmic¹⁵ like film, but are functions which depend more on environment than pre-programmed circuitry in the nervous system. Our reaction to light and heat, taste and touch and so on, are all very complex functions; we don't even know what all the functions are because there is so little research done on them. For example, color response functions may be different than pure white light. Neurological methods can help capture those parameters.

Luminance, color and saturation — as linear functions — have defined most television systems for the past 50 years; they work fairly well if replication does not have to be precise or represent "presence," though a number of compromising tricks have to be exercised to even reach levels acceptable for a general entertainment audience. For example, since the full human spectral sensitivity cannot be captured or displayed by red, green and blue, much less linearly, the blue is subtracted from the red to make "normal" colors like flesh tones look right to a non-critical audience. (Of course, this depends upon what one defines as flesh tones.)

Since those three factors are not sufficient to represent all of the information that the human visual system processes, essentially, what you get in the camera is not what you see, and hence the pattern that the photons from the display make on your retina is not necessarily what you see either.

That the simplistic model for vision works at all to produce acceptable if not perfect color television for most viewers is a miracle of cascaded engineering compromise, with the human eye adjusting for the missing characteristics not registered (or deleted) by the electronic system. However, the compromises made for television systems are not desirable for critical visual analysis, and may even be detrimental. The picture might be pleasing, and the motion might be satisfactory, but the output is just the opposite of what a scientific, medical, or even an artistic analysis of the scene requires.

[5.0] Some Surprising Research Results About Vision

Current research in the neurological and biological sciences have produced a host of unexpected and counter-intuitive data, essential to the design of next-generation motion imaging devices. There have been many surprises. For example, here are three major paradigm shifts in vision science which have been revealed in recent neurological experiments:

- Compression algorithms assume that motion vectors are the most important factor in an imaging system, yet recent data shows that the human vision system sees color changes several frames before an object's edge is detected and several seconds before the object itself would be perceived by the cortex. Indeed, there is evidence that if the brain does not integrate the visual perceptual process in this sequence — color, edge, motion — it will not have sensed the object at all (if it disappears from the screen), even though the retina has collected photons from the object and the optic nerve has transmitted the impulses to the cortex. This has immense implications for the inability of current systems to aid in scientific analysis.
- The process which causes our visual system to concentrate on selected areas in a scene had been assumed to be based on conscious visual cues. However, data indicates that *subliminal* cues are more important. Conventional video compression systems for entertainment physically suppress cortical attention mechanisms in the critical areas of a scene by removing such cues, just the opposite of what an analytic or "presence" system would require.
- The neural architecture for visual perception is organized and plotted in a hierarchy, from localized and specialized cells in the primary visual cortex that detect extremely simplified elements, to the more complicated processes responsible for complex perception. Research now shows the surprising result that the higher level combinatory mechanisms that humans use to process sensory input are located solely in the right hemisphere, outside of the visual cortex. This asymmetry has strong implications for image presentation, particularly for understanding depth perception which argues against a pure stereopsis¹⁶ explanation.

[6.0] Conclusion

It is now possible to build a body of information to construct new applied models of how we see. Advanced techniques have been developed using non-invasive cortical and computational methods, using tools such as functional MRIs, for exploring the relationships between human behavior or perception, and neurological processes, and to validate older psychophysical data. Older research methods, using electrophysiological techniques, with rare exceptions (e.g., during open brain surgery) have been confined to non-human primates, while other brain neuro-sensing devices, being non-invasive, can easily be used on humans with no ill effects. There is strong evidence that animals do not all perceive alike, so data collected from animals other than humans may not be relevant to designing next-generation imaging equipment. Such data has to be considered carefully in context.

The recent literature based on direct brain response has suggested important revisions to long-held theories on contrast detection, temporal and spatial color perception, motion sensitivity, object and pattern detection (especially of faces), spatial and temporal attention mediation, visual illusions, and the sequencing of subliminal and conscious events. Neurological data has begun to unravel these processes, producing a host of unexpected and counter-intuitive data which are essential to the design of next-generation electronic imaging systems.

These new results have significant implications for critical applications that rely on accurate image processing and visual perception, such as "telepresence" systems for remote visual scientific, medical, defense and engineering applications, and for archival record-keeping of unique objects, such as artworks.

[6.1] Proposed Research

Central to any data collection for advanced motion imaging systems must be an examination of relevant neurological processes using direct cortical access and similar analytic tools, and then to compare this neurological data against the conventional received wisdom based on older, less accurate and less perceptive psychophysical research. The main differences in these scientific approaches is that the latter technique merely observes reactions to stimuli, usually within constricted experiments, while the former seeks to understand cortical processes and interactions between and among neurological inputs. Therefore, data from psychophysics that is applied to imaging technology is often misleading. Moreover, psychophysics may mask

important underlying processes necessary for optimally matching equipment design to human perception.

We need to rectify this deficiency in imaging device design by first testing and validating the emerging neurological theses of human visual perceptual processing, and then establishing new design parameters for novel high-performance imaging devices coupled to next-generation transmission and computational systems. The following steps are then proposed after this novel data collection:

1. Investigate capture and display technologies which can exceed conventional ranges of color and luminance, and temporal processing;
2. Apply computational techniques to capture surface and sub-surface information across a broad spectral range (including ultraviolet and infrared), and with flexible and narrow spectral increments;
3. Tune the design and characteristics of capture, processing and display devices to the capabilities of the human visual system, as determined by the new human vision system models; and,
4. Exploit fully the capabilities of advanced transmission and processing systems, in order to replicate and enhance original scenes. It is no longer necessary to design imaging systems to be constrained by the limited bandwidth of older transmission systems.

These devices have numerous and obvious applications; however, we expect that novel insights may also emerge from the investigations of the human vision system, to wit:

- The human vision system has a known ability to distinguish adjacent points well beyond what conventional optical physical theory predicts is possible with our lenses and retinal structure. This is known as hyper-acutance; saccadic motion contributes to this ability, but in conjunction with cortical processes still poorly understood. Not only may it be possible to apply these techniques to imaging and processing (which is partly done today for specialized applications) but a better understanding of hyper-acutance may permit the design of end-to-end systems that further enhance the human perceptual process in this regard.
- Hyper-acutance, as noted, may be related to hyper-spectral cortical processes. If it is proven that color sensitivity is different for males and females, and is variable across the population (since color sensitivity is genetically linked), display devices should be uniquely configured for individuals to maximize and enhance information transfer. Electronically-stimulated hyper-acutance, in phase with our saccadic motions and spectral sensitivities, would be an extremely valuable mechanism for augmented vision.

- An investigation of how humans perceive motion may lead to signal processing of long-term rhythms in nature — the ability to discern behavioral patterns of the biosphere, for example. Motion picture systems have often shown such patterns if the temporal behavior (i.e., motion) is either slowed or speeded up, but normally only a human can “see” these complex patterns. A better understanding of how the cortex processes visual rhythms may lead to computer programs can emulate what we do in alternate time-space such as on other planets, and enhance our analytic ability to recognize useful life process indicators. This might be critical to the Mars mission, as well as for environmental analysis here on Earth.

[7.0] About the Author: Richard Jay Solomon

Richard J. Solomon is Senior Scientist, Program on Vision Science & Advanced Networking at the University of Pennsylvania's Center for Communications and Information Science and Policy. He is currently working in collaboration with Creative Technology LLC on the interfaces between super high-speed networking, electronic imaging and the human perceptual system for augmented cognition, sponsored by the Defense Advanced Research Projects Agency and the Office of Naval Research. This is a follow-up from exploratory projects at Penn since 1997 on the human visual system, sponsored by the Naval Research Laboratory and the National Security Agency.

From 1990-1997 he was an Associate Director of the Research Program on Communications Policy at the Massachusetts Institute of Technology. He joined the MIT program in 1977, and held a joint appointment as a Visiting Scientist at the MIT Research Laboratory of Electronics from 1989. As a principal in RPCP, he was instrumental in the creation of the MIT/Polaroid/Philips 720P/60 super high-resolution video camera, which demonstrated the results of psychophysical research begun in the 1980s at the MIT Media Lab, with which he was associated from 1986-90. He joined MIT in 1969 as a Research Associate in the Urban Systems Lab. From 1976-1980 he was also a Fellow at Harvard University with the Program on Information Resources Policy, where he researched regulatory and technology issues in telecom and transportation.

He is the co-author of *The Gordian Knot: Gridlock on the Information Highway* (MIT Press 1997), and numerous papers and books on telecom technology, regulation and transport. He holds three patents on Internet and telephone interfaces with other patents pending. He is currently working on a book on advanced technology for replicating visual presence.

[8.0] Endnotes, Bibliography and Further Readings

Endnotes

¹ Early psychophysical tests would have a subject look at a light dot, for example, and determine how dim it could be before he could no longer see it. Other tests measured reaction to color. These tests only measured correlation between stimuli and reactions, they did not address causation.

² More on functional MRI, <http://www.neuroguide.com/gregg.html>. More on PET, <http://www.epub.org.br/cm/n01/pet/petworks.htm>

³ Impedance is the degree to which an electronic component impedes the flow of current; the term is often used as “impedance mismatch” — a nerdy metaphor for what we might call “apples and oranges”, or two people (or systems) that have too little in common to communicate with each other.

⁴ Scan lines — the black lines in the picture similar to looking through window blinds — are an artifact of standard, “NTSC” television, which has been used since the inception of television. They are the areas the eye can detect which contain no image information. An NTSC signal uses an “interlaced” technique of breaking each frame image (480 viewable lines) into two sections called fields (240 viewable alternating lines). Because this process is happening quickly the eye is tricked into seeing the two fields as a full frame picture, albeit with flickering between the fields’s lines, aggravated further with color. This technique was acceptable on a size and type of set designed primarily for its day (in the 1940s a 10-inch diagonal black-and-white TV was considered large), NTSC images start to degrade rapidly as the screen size increases dramatically. And the introduction of color made degradation even worse.

⁵ An imaging system is how we mechanically capture light to reproduce it. We take something that is sensitive to light — a photosensitive electrical or chemical device, which could be traditional film, or a chip in a camera — and break up the image into little pieces, measuring, recording and storing information about the light intensity and color. If you look closely enough, or magnify any image enough, you will see the dots (like the “grains” in film, or the pixels on your computer screen) or lines (like the ones you see on your television) that comprise the picture. These mechanical means of breaking up an image into small components are what allows it to be displayed on a screen.

⁶ For a primer of how television works, see <http://www.howstuffworks.com/tv1.htm>.

⁷ A transmission channel could be a fiber optic cable, a wire that carries an electrical signal, or a radio wave carrying the signal of a picture from a transmitter to a receiving antenna. They are expensive because commercial broadcasters, such as television network operators, are forced by the economics of their business to transmit pictures at the lowest possible cost, maximizing profit by maximizing resource utilization.

⁸ Unwanted electrical or electromagnetic energy that degrades the visual or audio quality of the signal.

⁹ Technically, what we designed was a 720-line vertical and progressive (non-interlaced) coherent (non-scanning) raster transmitting at 60 frames-per-second.

¹⁰ *The Computational Brain*, MIT Press, 1992

¹¹ <http://www.gcsec.org/overviewshimmer.htm>

¹² Bandwidth is the amount of information that can be transmitted over a medium at any given time. In analog transmission systems, it is usually measured in megahertz; in digital systems, in bits per second.

¹³ For an in-depth discussion, see Neuman, McKnight & Solomon, *The Gordian Knot: Political Gridlock on the Information Highway*, MIT Press, 1997.

¹⁴ A trial in which neither the experimenter nor the subject knows which of several possible variables the subject is experiencing.

¹⁵ Logarithmic functions are a more accurate mathematical description of the luminance information. Film responds logarithmically while conventional television is designed to respond linearly to luminance and color. However, humans use neither function, but respond closer to exponential functions which are extremely complex.

¹⁶ Also known as “stereoscopic vision,” stereopsis is the triangulating function performed by the visual system that performs depth perception.

[8.0] Bibliography and Further Readings

Akins, *Perception*, New York: Oxford University Press, 1996.

Barten, *Contrast sensitivity of the human eye and its effects on image quality*, Bellingham, WA: SPIE Optical Engineering Press, 1999.

Beck, *Surface color perception*, Ithaca [N.Y.]: Cornell University Press, 1972.

Benson and Whitaker, *Standard handbook of video and television engineering*, New York: McGraw-Hill, 2000.

Berthoz, *The brain's sense of movement*, Cambridge, MA: Harvard University Press, 2000.

Boring, *Sensation and perception in the history of experimental psychology*, New York: Irvington Publishers, 1977 (1942).

Charnwood, *An Essay on Binocular Vision*, New York: Hafner Publishing, 1950 (1965).

Chevreul and Birren, *The principles of harmony and contrast of colors and their applications to the arts*, New York: Reinhold Pub. Corp., 1967 (1854).

Churchland and Sejnowski, *The computational brain*, Cambridge, MA: MIT Press, 1992.

Cornsweet, *Visual perception*, New York: Academic Press, 1970.

Davidoff, *Differences in visual perception: the individual eye*, London: Crosby Lockwood Staples, 1975

Dennett, *Consciousness explained*, Boston: Little Brown and Co., 1991.

Ditchburn, *Eye-movements and visual perception*, Oxford: Clarendon Press, 1973.

Dowling, *The retina: an approachable part of the brain*, Cambridge, Mass.: Belknap Press of Harvard University Press, 1987.

Evans, *The perception of color*, New York: Wiley, 1974.

Evans, *Eye, film, and camera in color photography*, New York: Wiley, 1959.

Farah, *The cognitive neuroscience of vision*, Malden, Mass.: Blackwell Publishers, 2000.

Gegenfurtner and Sharpe, *Color vision: from genes to perception*, Cambridge; New York: Cambridge University Press, 1999.

Gregory and Gombrich, *Illusion in nature and art*, London: Duckworth, 1973.

Gregory, *Eye and brain: the psychology of seeing*, Princeton, N.J.: Princeton University Press, 1990.

Hubel, *Eye, brain, and vision*, New York: Scientific American Library: Distributed by W.H. Freeman, 1988.

Hunt, *The reproduction of colour*, Kingston-upon-Thames, England: Fountain Press, 1995.

Hurvich, *Color vision*, Sunderland, Mass.: Sinauer Associates, 1981.

Jameson, et al., *Visual psychophysics*, Berlin, New York,: Springer-Verlag, 1972.

James, *The theory of the photographic process*, New York,: Macmillan, 1977.

Julesz, *Foundations of cyclopean perception*, Chicago: Univ. of Chicago Press, 1971.

Judd and Wyszecki, *Color in business, science, and industry*, New York: Wiley, 1975.

Katz, et al., *The world of colour*, London: K. Paul Trench Trubner, 1935.

Kolers, *Aspects of motion perception*, Oxford, New York: Pergamon Press, 1972.

Kosslyn and Sosherson, *Visual Cognition*, Cambridge, MA: MIT Press, 1995.

Land, et al., *Edwin H. Land's essays*, Springfield, VA: Society for Imaging Science and Technology, 1993.

Le Grand, *Light, colour, and vision*, New York: Wiley, 1957.

Le Grand, *Form and space vision*, Bloomington: Indiana University Press, 1967.

Lipton, *Foundations of the Stereo-scopical Cinema: A Study in Depth*, New York: Van Nostrand, 1982.

Lipton and Roaman, *Lipton on filmmaking*, New York: Simon and Schuster, 1979.

Lynch and Livingston, *Color and light in nature*, Cambridge, UK; New York: Cambridge University Press, 2001.

Luckiesh, *Visual illusions; their causes, characteristics, and applications*, New York,: Dover Publications, 1965.

Marr, *Vision: a computational investigation into the human representation and processing of visual information*, San Francisco: W.H. Freeman, 1982.

Minnaert and Seymour, *Light and color in the outdoors*, New York: Springer-Verlag, 1993.

Mollon and Sharpe, *Colour vision: physiology and psychophysics*, London; New York: Academic Press, 1983.

Nørretranders, *The user illusion: cutting consciousness down to size*, New York: Viking, 1998.

Ottoson and Zeki, *Central and peripheral mechanisms of colour vision: proceedings of an international symposium held at the Wenner-Gren Center Stockholm, June 14-15, 1984*, Basingstoke Hampshire: Macmillan, 1985.

Palmer, *Vision science: photons to phenomenology*, Cambridge, Mass.: MIT Press, 1999.

Papathomas, *Early vision and beyond*, Cambridge, Mass.: MIT Press, 1995.

Pickford, *Individual differences in colour vision*, London: Routledge and Paul, 1951.

Purves, *Neuroscience*, Sunderland, Mass.: Sinauer Associates, 1997.

Ratliff, *Mach bands: quantitative studies on neural networks in the retina*, San Francisco.: Holden-Day, 1965.

Rock, *Perception*, New York: Scientific American Library: Distributed by W.H. Freeman, 1984.

Ronchi, *Optics, the science of vision*, New York: New York University Press, 1957 (1991).

Ross, *Behaviour and perception in strange environments*, London: Allen & Unwin, 1974.

Slyce and Hughes, *Patrick Hughes: perverspective*, London: Momentum, 1998.

Stein and Meredith, *The merging of the senses*, Cambridge, Mass.: MIT Press, 1993.

Uttal, *The psychobiology of sensory coding*, New York: Harper & Row, 1973.

Uttal, *A taxonomy of visual processes*, Hillsdale, N.J.: L. Erlbaum Associates, 1981.

Stevens and Atkinson, *Stevens' handbook of experimental psychology*, New York: Wiley, 1950 (2002).

Wade, *The art and science of visual illusions*, London; Boston: Routledge & Kegan Paul, 1982.

Watkinson, *The art of digital video*, Boston, MA: Focal Press, 2000.

Watkinson, *The MPEG handbook: MPEG-1, MPEG-2, MPEG-4*, Oxford [England]; Boston: Focal Press, 2001.

Williams, *Image clarity: high-resolution photography*, Boston: Focal Press, 1990

Wyszecki and Stiles, *Color science: concepts and methods, quantitative data and formulae*, New York: Wiley, 1982.

Zeki, *A vision of the brain*, Oxford; Boston: Blackwell Scientific Publications, 1993.

Zeki, *Inner vision: an exploration of art and the brain*, Oxford ; New York: Oxford University Press, 1999.