A Taxonomy of Natural Vision Systems

Brian McMillin

Abstract

Vision Systems have evolved in many different configurations in the natural world. As species adapt to their environment they adopt different strategies to balance the needs of resource management, feeding, reproduction and protection from predation. In many species an effective vision system is a dominant element of this strategy. Interestingly, the major components of a vision system involve many aspects of the organism that are unrelated to the actual detection of light.

Frequently, certain standard physiological structures within the organism are given new capabilities through the

- coordination of diverse structural elements
- new usage strategies
- evolutionary adaptation into new structures.

This Taxonomy attempts to list individual features and strategies, recognizing the species and groups that form a linked evolutionary tree as well as those that appear as independent developments.

In addition, the evolution of features and strategies as coordinated groups across species highlight their importance.

Contents

Background

Vision systems first evolved as a consequence of photosynthesis. Techniques to optimize exposure to light enhanced the survival of organisms with this capability. These techniques included the ability to detect directionality and strength of light, as well as the ability to respond by modifying the organism's movement patterns.

Many plants evolved a growth strategy that incorporates multi-dimensional branching. This enables effective coverage of an energy-rich or nutrient-rich environment, as in the case of stems, branches, leaves and roots. It also supports wide dispersal of seeds or spores. However, this organization becomes too unwieldy to support a nervous system.

As a survival strategy, bilateral symmetry is one of the most prevalent and effective. Given pairs of organs, species are free to try various adaptations within the general design. For example, humans generally coordinate the use of their eyes but one eye is given a dominant role. Some species achieve heightened acuity by increasing coverage of a target area, while others optimize coverage area by using non-overlapping fields of vision.

A common boundary between the two symmetric organs simplifies the placement of the nervous system. This provides a more protected location even before the development of advanced features like a spinal column.

Redundant organs provide enhanced survivability. This is true in the obvious case of damage incurred during the course of the organism's life, but probably more importantly in the case of potential developmental defects during ontogeny. Increasingly complex organisms cannot be reliant on perfect growth and development at every stage.

High-level redundancy helps to prevent an unacceptable rate of fatal defects during early development. The visual systems under consideration here tend to work best with vast numbers of repeated low-level structures such as photoreceptors. It is unreasonable to expect perfect regularity or identical development within a single organism, or from one organism to another. The addition of more repeated interconnecting structures, such as neurons, with an

essentially random *initial* structure, solves the development problem and increases the functionality of the entire system by *learning* to deal with this randomness.

The ability of neural development to empirically adapt to the actual structures that grew in a particular individual adds flexibility and speed to the growth of the individual and, thus, the entire species. Maximal fault tolerance. Defects such as color-blindness present only minor impairment to the individual. Enhancements such as tetra-chromaticity can be tried on an individual basis without overt risk to the individual or species. Mutations that relate to these changes can be implemented as a measured response, allowing gradual adaptation to environmental niches. This eliminates the go/no-go, live or die, binary response to mutations and allows large complex organisms to experiment with thousands of potential changes simultaneously. Evolution approximates the solution to the thousand-variable min-max problem in a small number of generations - with generally non-fatal results to the individuals involved.

Even given an underlying bilateral symmetry, the target environment and life style have led flatfish to place both eyes on the same side of their body. This radical change in shape and position of the two “developmental sides” is only one example. In humans, one lung has two lobes, the other has three. Really complex plumbing changes allow one side of the heart to exclusively service the lungs, while the other side services the entire rest of the body. And viable individuals exist wherein the structures developed in mirror-image from the “typical” locations.

Compound eyes allow an evolutionarily balance of complex, replicated fixed structure vs. the heightened energy requirements of active musculature and neurological control elements. Remember that even “less advanced” organisms such as arthropods are actually solving highly complex problems and are achieving maximum efficiency when adapted to their particular environmental niche. In addition, creatures with compound eyes have had the opportunity to experiment with many light-gathering and light-focussing techniques. Some of these include lenses, prisms, flat reflectors, corner reflectors, tubes, masked edges and filters. In some cases polarization can be sensed.

Common problems, such as trying to achieve maximum resolution as well as maximum field-of-view often find similar solutions in unexpected places. The fovea in a human eye uses a high-resolution central region, with progressively poorer resolution toward the periphery. Flies, bees and mantises also use a fovea, but one that is created geometrically by flattening part of the compound eye to make more sensors cover a smaller angular region.

Physiological Structures

Light Receptors

Data Processing Structures - Retina and Optic Nerve

Eyeball

Compound Eye

Musculature

Eye Location

Optical Structures

Light Receptors - number of elements, angular density, percent coverage, effectiveness
    Humans up to ~150,000 per mm^2 toward central Macula

Filters and colored oils

Fovea / Macula - size and sensor density

Blind spot (optic disc) - location, limitations and compensation
    All vertebrates have a blind spot, all cephalopods do not
    Position in humans 12-15 deg temporally and 1.5 deg below horizontal
    Size in humans 7.5 deg high and 5.5 wide

Optic Nerve - Vertebrate neurons on front of retina

Cornea

Lens

Iris

Focus - Range of accommodation

Reflective retinas (tapeta lucida) dual-pass photon absorption, nocturnal predators and pelagic

Visual Capabilities

Spectral Response

Dynamic Range in brightness and contrast

Ability to detect polarization

Sensitivity Range

Response Rate

Motion Sensitivity

Motion Tracking and precision timing

Cancellation of observer motion and artifacts like blinking

Field of Vision

Spatial Resolution

Axis Alignment (targeting)

Focal Range and Response

Depth of Field

Iris Shape and Responsiveness

Situational Acuity

Envision a graphical presentation with a spectrum of situations on one axis and each of the capabilities on the other axis. The quality of each capability could be presented as a color.

This graphical rendering allowing various species to be compared and contrasted.

TBD - Create example template and replicate throughout this document as active examples

Variations Within Individual Species

Gender-Linked Variability
- Prevalence of Color-Blindness
- Prevalence of Tetra-chromaticity

Eye Color and Environment

Interspecies Visual Adaptations

Examples include the adaptations of prey to incorporate camouflage in color, shapes or structure to reduce the effectiveness of a predator's visual system.

Less obvious are adaptations that trigger innate aspects of another species' visual capabilities. Examples would include bright coloration of poisonous animals. The colorful adaptation evolved in coordination with the predator's hard-wired responses and need not involve individual learning - as in the case of fatal poisons.

This evolution within the prey is only beneficial when the aversion is present in the primary predator or group of predators. It must not be outweighed by other aspects of increased visibility or the costs of creating or maintaining the coloration.

Intraspecies Visual Adaptations

Frequently multiple concurrent adaptations within a species lead to enhanced survival.

The combination of simple, hard-wired visual circular-target-shape detection and the presence of areoles enable human infants to receive more nutrition with less energy expenditure on the part of both mother and child. Douglas Hofstadter's hypothetical Grandmother Cell (a neuron that fires whenever a representation of your grandmother is seen) can be replaced with a more concrete, survival-oriented example where the DinnerTime neuron fires when the mother's breast is seen.

Another commonly mentioned adaptation is the gender-based difference in coloration seen in birds. The bright coloration of males triggers a mating response in females, but an aggression in other males. Likewise the dull female coloration is related to camouflage; bright male coloration tends to distract predators.

Periodic Variations

Diurnal Changes

Seasonal Changes

Developmental Phases - Larvae vs. Adult

Developmental Changes - Infancy, Childhood, Adulthood

Age-related changes and visual longevity - presbyopia

Environmental Considerations

Air / Water

Day / Night / Underground / Deep Sea

Operating Temperature

Humidity

Thermoregulation requirements

Diet and Raw Materials

Calorie Requirements - system energy utilization

Moisture

Proteins and Amino Acids for growth, regeneration and maintenance

Vitamins - requirements and deficiencies

Poisons or other adverse materials

Nature vs. Nurture

Consider schooling of newly-hatched fish that could not have learned from behavior of a previous generation. Inherent responses. Rapidly learned responses: add a naive fish to an established school.

Operational Goals

Finding Food

Finding Shelter

Avoiding Threats or Predation

Finding Mates

Finding Home

Friend / Foe recognition - mates, family, clan, other

Coordinating Attacks

Coordinating Pack / Flock / School

Motion Tracking - Hitting a fastball

Visual Development and Training

The process of developing an effective visual system within an organism involves a sometimes lengthy learning process. The ability to use multiple structures in a coordinated manner requires experience with stimulus-response patterns and reinforcement of appropriate neural pathways. Human newborns must learn to focus their eyes over a period of days and then move on to motion tracking involving the coordination of both eyes.

Experience with visual patterns within an environment will, over time, enable a categorization of low-priority, familiar features and new, unique and interesting high-priority features. The ability to identify and prioritize new objects allows higher-level resources to be allocated to learning and efficiently building a world-view incorporating new or infrequent experiences. The development of territorial familiarity enhances the opportunities to roam for food and mates, to return to home, nest or burrow, and to avoid danger in the process.

The size of the available memory, the richness of the visual environment and the allowed developmental time all play a part in the strategies used by particular organisms. Some species have visual pathways only a few neurons long that are ready to function effectively immediately after hatching. Humans spend years developing the skills to allow visual coordination of balance that enables walking upright. It takes even longer to develop the hand-eye coordination used to throw or catch a ball - or to use a computer mouse.

The time required for this neural development implies a delay before individuals can be evaluated for relative fitness. The qualities (or lack thereof) of individual structures, and the learned abilities associated with them, are not immediately obvious. This leads to an extended childhood period. The entry into adulthood begins the series of reproductive choices that select traits for the next generation.

The quality of the visual system, the size of the supporting brain structures and the social structure of the species all interact to inform the capabilities of the visual system. Truly advanced visual capabilities require long periods of training that can only be available in a supporting society that affords individuals sufficient time periods free of the risk of misadventure or starvation.

Just as learned motor skills require less and less attention with practice and repetition. Walking across the floor, for example, is accomplished without thought. Subtle guidance from multiple sources causes adaptation of speed, gait, foot placement, etc., also without conscious intervention.

Common vision skills also become more and more “autonomic” with practice. Walking down a corridor involves observations of the entire visual field as it changes over time. Recognition of the rate at which a wall is approaching as compared to the rate of forward motion happens at an extremely low level. This results in issuing simple, very subtle suggestions to the motor

system. Learned vision skills such as these are pushed further and further from conscious thought.

The ability to walk through a doorway is so routinely successful that no consciousness is involved. Passing an obstruction in the doorway, however, may rouse some level of consciousness and invoke higher route-planning to provide speed or timing adjustments. Only a total blockage of the path will thwart the unconscious ability to perform the overall, complex sequence of movements.

Avoiding running into objects is a visual skill that does not require any complex object recognition or higher-level functions at all. It is only necessary to discern relative motion over short time spans in various parts of the visual field. Once in motion, the relative changes in the scene form the sensory part of a simple feedback system that makes small adjustments to the muscle-memory involved in walking. This makes the process very fast and resource-efficient. The automatic nature of these casual, repetitive operations means that they do not interfere with simultaneous, higher-level functions such as threat assessment.

Facial Recognition

The general concept of a face is nearly universal among mammals. Birds and reptiles may also have face-like structures. Bilateral symmetry generally causes two eyes to be located in proximity. It is nice to be able to see what you are eating, so a mouth near the eyes would be helpful. Sensors for smell and taste are tightly coupled, and it is a very good idea to make sure food smells right before eating it. Thus, eyes, nose and mouth will tend to evolve as a group.

The face-like structure detection has evolved to be very fast. In general, the alarm afforded by a fleeting glimpse of a saber tooth cat in the bushes likely allows you and your offspring to survive; failure to notice that pattern: not so much. Seeing faces in ordinary objects (pareidolia) - unless carried to extremes - is thus a survival trait. Better safe than sorry.

Initial detection of a face-like structure can initiate a fight-flee-or-freeze decision tree. When additional cognitive effort is required the decision to freeze is usually the safest. If it is a predator, freezing will not draw undue attention. If it is an enemy, freezing will not be perceived as a premature attack. In some species the startle response is to jump, and the predators have adapted to expect that jump.

The face-seeing survival trait is also the genesis of art. A casual sketch can convey specific meaning to others without requiring precision or sophisticated tools. The meaning may include individual or group identity, emotion, and intent. All are attributes that we readily recognize in familiar faces.

Visually recognizing faces would be a survival trait for prey, and developing recognition-countermeasures such as camouflage would tend to favor predators.

Facial recognition skills lie in a continuum and vary in their speed and computational sophistication. Examples include:

- Is it a face or not?
- Is it my kind of face? i.e. family or friends
- Is it dangerous? Or tasty?
- Is it looking at me?
- Is it a specific individual?
- Is it trying to tell me something?

Often facial recognition is used only to select a rapidly-identifiable starting point. The face is an anchor for broadening examination of the visual field. Individuals who are partially obscured tend to be easier to identify if all or part of their face is visible.

The speed, accuracy and volume of data generated by the vision system will be balanced in evolutionary terms against the needs of the organism. Highly specialized (possibly genetically ordained, hereditary) features may be perfectly adequate for one species in a particular environment. More generalized capabilities, possibly with the capability of long-term learning and adaptation, may be more effective for a different species' needs.

Recognition of *individual* faces is an extraordinarily difficult task. The range of expressions that are so useful for nuanced communication in a one-on-one situation run counter to casual recognition in everyday situations. All manner of clues that have nothing to do with *actual* facial recognition are used to enable rapid identification of individuals. Patty is the one with triangular earrings and Selma's are round - but the faces themselves are literally identical. The context allows our brain to casually winnow out the unlikely candidates: you fail to recognize your grade-school teacher in the check-out line at the grocery store. Even given accurate context, I can still (embarrassingly) fail to recognize my short-haired literature teacher when she comes to class wearing a wig.

Human beings spend years and tens of thousands of hours training their brains to successfully recognize the people they associate with in all situations, angles, distances, lighting, moods and activities. Familiarity continuously updates the associated neural nets to maintain recognition accuracy. Even so, simple lapses such as a week's vacation in the sun, a new hair style or different glasses can cause a noticeable delay in recognition. In addition there will be a significant amount of additional study ("Wow! Let me get a look at you.") while critical features are selected and added to the training data.

This critical primary training occurs in childhood. The brain learns to recognize facial features and creates groups of features that are either distinguishing characteristics or noise (irrelevant). This distinction may change with the situation - squinting means one thing indoors and another in bright sunlight.

This key recognition ability enables the recognition of subtle identifying characteristics from within the clade and clan of individuals that form the training set. This ability works well enough for normal social interaction but progressively breaks down as situations become less familiar. Identifying people in a family photo album or school yearbook can actually be an intellectual challenge.

The choice of characteristic markers is highly tuned to the needs of recognizing individuals within a clan. Encountering an individual from outside the clan is (by definition) rare and is not a well-honed skill. Initially, it is sufficient to classify them as "other" and be done with it. It requires multiple, repeated exposure to individuals from outside the clan to allow the brain to begin to do its job and start assigning unique markers to these outsiders. Without sufficient exposure, the racist cliche "they all look the same to me" is literally true. Your brain has insufficient information to distinguish unfamiliar individuals.

This entire aspect of the operation of the brain means that any "other" individuals will not only fail to be rapidly recognized, but facial expressions and emotional nuance will also be missing. The inability to do the basic, casual recognition will cause an increased conscious workload and lead to a higher level of situational anxiety. The only way to combat this prejudicial bias is to ensure the most cosmopolitan environment possible, especially during early development of the brain. Ensuring the ability to recognize the importance of different facial markers essentially

expands the definition of clan - even if the family group is still predominant. Individuals must be seen in natural situations displaying a range of gestures and emotions. Television (and more modern technologies) can be very beneficial in this respect, as long as intentional efforts are made toward diversity. Casting John Wayne as Genghis Khan does not count.

Recognizing these inherent limitations in the visual system, and the types of coordinated "tricks" that we routinely use to make the system work better, leads me to seriously question the usefulness of things like mug books and police line-ups. The expectation of any accuracy in comparing a standardized photo in an album with the memory of a glimpse in a dark alley during a mugging is outlandish. Further, since experience shows that this mug shot approach does not actually work in the general case, there is a strong temptation toward selection bias: to "help it along" by only presenting the album with the "skinny black guys". Plus the obvious issue that the photos had to come from somewhere (i.e. previous offenders) and are of unknown age (they certainly are not current). This entire process is also critically dependent on the ability of an average individual to extrapolate facial similarities between a standard photo of unknown age and the fleeting memory of a traumatic situation. Bottom line: eye-witness identification is provably unreliable.

Machine learning systems attempt to do facial recognition using several approaches. Most involve detecting certain standardized marker points in a photograph and measuring the geometry of these locations. When the face is seen again, a decision is made as to whether the geometry is "close enough" to the previous values. This requires all photographs to meet certain arbitrary standards, which are unlikely to occur in any natural situations.

The standardized marker locations used by most facial recognition systems are not a part of the machine learning / neural net portion of the system. They are arbitrarily selected, *a priori* points bestowed on the system by the (invariably white) system designers. The marker points were not learned by the system based on the importance to the recognition task. This decision is made in the anticipation of increased efficiency and the assumption that "of course the corners of the eyes and mouth are important locations". Unfortunately, these markers are not necessarily of any particular value to recognizing individual members of an arbitrary clade.

The promise of machine learning relies on the ability to deal with vast quantities of input data. One picture of each individual does not constitute *vast quantities* of data, and therefore cannot be expected to yield much more than random results.

The most reliable facial recognition is Apple's FaceID, which uses a 3D mapping and is expected to generate a Yes/No result based on a single individual. FaceID updates the internal neural net every time it is used. This creates a training database with a large number of samples of the same individual in a different poses and expressions taken over time. FaceID does not use visible-light photography and cannot be compared to any natural vision system. On the other hand, the neural network, continuous training, and the "is this who I think it is?" aspects do correlate to the operations of natural systems.

Any system that claims to be able to look up a specific individual from a directory of historical photographs is questionable. Comparing standardized images to photographs taken in a natural setting is unlikely to succeed. All that can be hoped for is a broad-stroke estimation that can be used to eliminate most candidates. The remaining cases must be carefully examined by trained individuals.

The facial markers required by many of the common facial recognition database systems are so incredibly touchy they border on being unusable. I personally had to try six consecutive times, in a controlled setting, to obtain a passport photograph that was deemed acceptable to the system. And the resulting, standardized photograph (stand on the line, white background, look straight ahead, neutral expression, no glasses) does not correspond to the way I would ever be seen in a natural setting. Furthermore, the photographer, who takes these pictures everyday, could not anticipate whether the system was going to accept or reject any particular picture.

Some facial recognition techniques appear to work well for helping to identify individuals in a photo album. In general, this uses a directory of a few dozen individuals and the system can make a fairly good guess as to which is which, even is natural settings. The actual number of candidates it selects from is surprisingly small (the mother is unlikely to be mistaken for the baby). If Aunt Matilda is manually tagged as being on the trip to Cancun, then chances are good that this other (barely identifiable) blob in a Cancun picture might also be Aunt Matilda. This is an example of "tricky outside information" that the system (and a human being) will routinely use to boost the odds of correct identification. These make for apparently impressive results. From a numerically objective standpoint they lose their glamour. And trying to scale up systems based on this type of technology is essentially hopeless.

I can rapidly and efficiently spot my wife in a crowd. This has little or nothing to do with actual facial recognition. The actual visual skill falls into the category of object recognition, and the rules are: Use any skill or knowledge you have to find the object that is my wife.

I have detailed recent knowledge of the characteristics of my wife: height, mannerisms, hair style and clothing, viewed from many previous aspects. I can move through the crowd, changing my viewpoint and gathering additional, current information. The goal is to pick out even the tiniest glimpse of any of these wife-aspects and see if they could possibly represent a piece of the larger whole. There may be dozens or hundreds of false starts, but finally a momentary flash of a particular sneaker or scrunchie resolves itself into the desired target.

Our high-performance natural vision system is capable of routinely solving problems in this style. The requirements are continuous, rapid matching of small portions of the visual field against memories of similar objects and their previous surroundings.

It is possible to train any recognition system (natural or artificial) to perform well with any given set of input data. Those same systems will fail miserably when presented with inputs that have

no exemplars in the training. It is incumbent on the trainers to present the widest possible range of experience. Any recognition task must be able to generate a result of "I don't know". Forcing some arbitrary "choose the best match" requirement is guaranteed to create erroneous results - especially in the (inevitable) case of encountering novel input data.

The arbitrarily selected facial markers mentioned above are also used for creating artificial faces in animations and deep-fakes.

In the case of deep-fakes, a training set of short video clips are run through a recognizer to analyze the geometry and geometry-change sequences that are characteristic of the particular individual's expressions and speech mannerisms. This seems like a clever idea, but it falls short upon further examination. The training inadequacies, primarily exemplified by the arbitrary choice of marker locations, are magnified by the double usage. Essentially, the mistakes on input are then multiplied by mistakes on output.

The casual observer of a deep-fake will tend to accept it at face value. A discriminating observer, especially one from the individual's own clan, will easily recognize the artificiality, even if he cannot articulate *why* the fake is wrong. The observer falls into an uncanny valley.

Wholly synthetic videos, for example ones using motion capture as input, can create acceptable faces and animations because disbelief has already been suspended. The knowledge of the artificiality lowers the expectations. Just as artificial landscapes will rarely fool a geologist or botanist, synthetic humans will not fit into any natural clan.

Subtle similarities are recognized routinely within clade and clan. Declaring "You have Uncle Harry's eyes" is not a statement that any synthetic recognition system is capable of making with any credibility.

There is no such thing as Speaker-Independent voice recognition. The illusion of speaker independence is achieved by learning thousands of different accents, dialects, ages, genders and colloquialisms in order to select a best-guess speaker-equivalent and narrow down the possible utterances. Learning the patterns of a specific individual is always preferable, and improves speed and accuracy. These statements apply to both natural and artificial recognition systems.

It should be realized that the visual systems must work the same way: clade-independent facial recognition does not exist. Many samples of different expressions and gestures in different situations must be experienced from each unique clade in order to have any expectation of accuracy in understanding the identity of the individual and the meaning he is communicating.

Just as people learn to understand different dialects and accents through hearing sample speech, they must be exposed different facial characteristics in natural activities. The pattern of facial expression changes is as important as the cadence, pitch and prosody is to human speech.

Speech recognition systems must make estimates that could be indicative of cultural background when identifying accents and dialect. Similarly, universal facial recognition systems will necessarily make estimates of ethnicity when narrowing down clade membership.

In order to avoid stereotypical behavior and racial, ethnic or cultural bias it is necessary to provide any recognition system with near-universal training materials. This means video and audio samples of as broad a spectrum of individuals as possible. Ensuring these samples are free of inherent bias and represent a full range of gestures and emotions is fundamentally challenging.

Providing such idealized training to artificial systems is technically possible, but reveals the true magnitude of the task. The fact that the magnitude is infeasible with today's systems simply makes the goal an aspirational one.

Providing similar idealized training in a human context is similarly aspirational. Such training would require the materials to be structured in an engaging, thoughtful manner. The general idea would be to provide students with a breadth of exposure to faces and speech that expand their clan and augment their brain's inherent recognition capabilities. The learning process must occur while the student is actually "paying attention" but it must not be the focus of the activity. The acquisition of nuanced communication skills is always a background task. The brain discovers unique correlations among sequences of events through repetition in dissimilar situations. Valid correlations are reinforced during immersion in other tasks.

Insufficiently diverse training results in brittle systems. And, by definition, that brittleness will reveal itself in unexpected failures.

MOTION BLUR AND OBJECT TRACKING

Light impinging on the retina is blurred by the motion of objects in the environment as well as by the motion of the eye itself. Any thing that can reduce the magnitude of either of these components will improve the clarity of the image.

If a particular object of interest can be tracked and kept nearly stationary in the visual field it becomes easier to resolve details and changes in that object. The coach reminds you to keep your eye on the ball.

If the motion of your head can be cancelled out, the visual tracking becomes easier. The vestibular system provides position and acceleration information, primarily to allow a mammal to maintain balance. The vestibulo-ocular reflex provides an extremely fast direct neural link that assists in image stabilization on the retina.

DETECTING ANOTHER ANIMAL'S ATTENTION

Stalking Prey while the Prey is not paying attention. Knowing when to strike so the prey's response will be delayed.

Cats that do the tail-twitching distraction technique, coupled with the visual skill of estimating the focus of the prey's attention, have an evolutionary advantage. It doesn't have to be much, and it certainly doesn't have to work every time.

IMPROVED RECOGNITION DEVICES

- Combine audio-video data acquisition from real clades worldwide incorporating a range of gestures and emotions.
- Use this information to enable nuanced motion capture and recognition. Use this to generate synthetic individuals in artificial situations.
- Use this to build movies or video games.
- Use this to create immersive training data that is responsive to individual students and situations.
- Use this to create artificial assistants that are responsive to gesture and emotion.
- Use this information to enable improved image recognition systems that more accurately focus on elements of nuance and exclude non-informative noise.

IMPROVED RENDERING DEVICES

Use this understanding to inform the creation of more accurate visual rendering or projection devices. Focus less on repetitive display of background pixels and more on portions of a visual field that are actively carrying information through the entire length of the viewer's processing chain. Use knowledge of that processing chain to emphasize areas and actively draw attention to aspects of a presentation. Conversely, draw attention away from unimportant or incompletely rendered areas.

Recognize that all modern display devices are based on technologies specifically tailored to trick the *human* visual system into into accepting an illusion of a possible reality. All aspects - brightness, contrast, color choices, resolution, frame rate, etc. - are chosen only with "average modern humans" in mind.

A domestic cat will sit at a window and watch the outside world for hours, but will show no interest in a television. The video image appears cartoonish as far as the cat's visual system is concerned, and therefore no information passed to its higher brain functions is sufficient to engage more than casual interest.

It is important to realize that the humans are voluntarily "suspending disbelief" when presented with synthetic images. They willingly allow imperfect representations to pass information through the visual system to the higher brain functions because that information is ultimately interesting - not because the representation is inherently accurate.

A van Gogh or Picasso painting, or a black and white movie, makes no attempt to accurately render a reality. Yet they are all engaging to the human brain and worthy of extended consideration.

The goal of future display technologies (and future media design) should be to convey the desired high-level message to the specific audience. This means that it is not necessary to have infinite resolution, multi-spectral displays that cover the entire visual field. Such things would be adapted to the capabilities of the human visual system - but the purpose of the human visual system is to *reduce the redundant information content* before passing the data to higher brain functions.

Key survival traits of the visual cortex are its ability to detect discontinuities, abnormalities or unexpected departures from experience in the visual scene and flag them for closer inspection. But these warning signs are being actively suppressed by the higher brain since it has agreed to suspend disbelief.

Therefore, all the effort put into creating images that are indistinguishable from reality is immediately discarded and the only information passed on from the visual cortex is what it deems interesting or significant - with no reality-warnings at all.

INTERACTIVE VISION

This understanding will be especially important for truly interactive visual displays. The display should react to, and anticipate, the particular viewer. The data feed for each viewer should be unique. Wandering attention should be detected and redirected. Performance quality should be enhanced by learning what is appealing to the viewer.

The set of responses from the viewer can be used to generate a profile indicative of their history and experiences. This can be used to fabricate a more meaningful or immersive experience. These adaptations can be used to generate more accurate communication.

Behavioral Anticipation

The ability to recognize particular patterns of behavior exhibited by other members of one's own species is a key aspect of visual communication.

Certain inherited behaviors carried out by one individual may require a learned response to be developed in others in its group. This visual adaptation will be generally more responsive to individual variations and nuance of combined behaviors.

The ability to signal intent to other members of a group is critical to pack hunting as well as establishing and enforcing an organizational hierarchy.

Subtle gestures and expressions all require a visual time sequence to be recognized. This can provide very rapid, accurate and silent communication.

Interactive gestures can ensure one-on-one understanding between individuals. More universal "broadcast" gestures can convey instructions to a group.

The overall visual system allows intentional communication through gestures in a particular situational environment. Symbolic language evolved to communicate similar combinations of intentions in particular situations. Representational art and written languages both perform their functions using exactly the same visual-sensory and neural hardware. The difference lies solely in the training of the neural network.

The predictive or anticipatory aspects of the neural network is what reinforces effective pathways. Neural networks are trained by taking a particular input and generating a set of hypothetical future states. Neural pathways that generate accurate predictions are enhanced whereas pathways that generate inaccurate predictions decrease in prominence. The choice of predictive timeframe is itself a parameter that may be learned by the network. Carried to an extreme, this predictive process is the basis of visualization of imaginary environments or situations.

Doing this communication with facial expressions is not always fast, accurate or efficient. Routine communication with a flick of the eye or nod of the head can be sufficient - especially when the immediate behavior of the other person confirms receipt of the message. At the other extreme, I could easily spend hours studying the Mona Lisa, building hypothetical worlds, imagining the thoughts that might have been going through her head, and weighing them against my own experiences. A recognition system with no experience would not even realize that there was anything interesting about her face at all.

Evolutionary pressure has provided impetus for greater capabilities in symbolic language among Humans. The visual recognition systems are generally common to all primates, but it is the ability to interpret observed symbols that is the differentiating factor.

The ability to use observed symbols to trigger the visualization of a wholly imaginary situation or action is what allows Humans to collect knowledge from each other across space and time. Written communication requires only a shared understanding of particular symbols. Sequences of symbols provide a recorded history of concepts that can be shared among any that share the requisite background. Thus, the learned knowledge acquired by any individual can greatly exceed the experiences of any single lifetime.

Visual Orienteering

Orienteering represents a strategy for navigation of both familiar and unfamiliar environments. This is critically important to safely allow foraging and timely return to a safe haven.

- Establish important reference points in the environment
- Maintain a direction heading by the alignment of distant objects
- Measure distance using the passage of near-field objects
- Identify turns by study of intersections for additional detail
- Allow for changes in lighting, such as different times of day or weather
- Routinely monitor the appearance of the back-trail to allow rapid return.
- Develop a notation to allow describing the path to others

Consider birds finding their own nests in trees

Consider migration

Consider ancillary data sources -
terrain and textures, energy expenditures, sounds, smells, magnetism

Consider rabbits locating their burrows

Consider bees finding flowers

Consider Boy Scouts or soldiers and their training

Consider the skills required to drive a car - anticipating lane changes and turns, returning home, compensating for road closures, etc. Consider the rapid skill deterioration of these skills caused by the introduction of GPS devices.

Other species will exhibit the same behavior - sharpening frequently used skills and letting others atrophy, even over comparatively short time spans, as the environment changes.

Note the critical nature of the ability to remember a temporal sequence of visual patterns in each of these examples.

Remember also that these visual patterns are only approximate - specific key features must be learned, recognized and matched even in the presence of changes of scale, orientation, lighting, etc.

Usage Strategies

Binocular Vision
Eye Dominance
Head and Body Movement
Saccades

VISUAL FIELD

Bilateral symmetry gifts the organism with two eyes, which allow it to enhance its performance. But an individual should always be able to survive with only one eye.

Clever herd animals are able to achieve a tremendous increase in visual coverage by almost eliminating overlap in the field of view of their two eyes. However, they do so in a herd environment where they share any alert of danger (or location of tasty grass) among the individuals. Thus, a one-eyed cow is not at a significant disadvantage.

Predators will, of necessity, attack straight ahead. This means that the redundant vision systems should be pointed in the direction of that attack. Again, the one-eyed individual is still functional and not fatally compromised.

DEPTH PERCEPTION AND RELATIVE DISTANCE ESTIMATION

Depth perception is a critically important feature of vision systems. A naive assumption may be made that stereo vision (implied only by the presence of two eyes) is the basis of depth perception. It turns out that this is not in any way required, and many techniques have evolved to enable single-eye depth perception.

Very near field depth perception can be achieved by using focal distance. The eye can adjust the focal length of the lens to achieve the brightest image of the target object, indicating that it is in focus. The accuracy of focal distance measurement may be improved by reducing the depth of field. Reducing the depth of field is achieved by increasing the aperture of the iris. Thus a cat that is stalking prey has big eyes.

The eyeball is not a single point. The diameter of the eyeball means that a significant parallax exists as the visual axis sweeps across an image. Near-field objects shift with respect to the background as the eye moves. Note that this parallax perception is achieved as a feature of the peripheral vision portion of the retina - it does not need the high-resolution fovea.

Additional single-eye parallax-based depth perception can be achieved by making use of nearby objects in the field of view. The parallax baseline is the shift caused by eyeball rotation, and a near-field object might be a branch or eyebrow whisker. The perceived shift of this nearby object can be compared to the shift seen in a target object during the same eye motion. Again, the relative distance is derived using peripheral vision and without the central vision field.

Many birds, for example have eyes on the sides of their heads to achieve maximum peripheral vision. They often (but not universally) evolve to use a head-bobbing motion to achieve depth perception. Head bobbing is only casually coordinated with a stepping gait, and is not actually associated with maintaining balance. The head-bobbing establishes a baseline for parallax to determine relative distances. In addition, the smooth bobbing motion provides the images with a continuum of parallax-derived relative motion, not just results from a two-image stereo pair.

Furthermore, the head-bobbing or head-rotation applies simultaneously to both eyes, thus giving the bird almost 360 degree relative-distance information. This near-field object distance information is important for tasks such as catching food.

Note that this head movement - increased parallax - distance estimation combination is routinely found in humans. Parking a car is considerably more difficult if you hold your head rigidly in place in the driver's seat. Your normal, casual head movement provides relative distance via the parallax of looking past the hood, fenders, door posts, etc. The head changes position with respect to these obstructions . Changes in visibility of distant objects provide a wealth of information to aid in accurately positioning the vehicle.

It is true that two eyes increase visual acuity and responsiveness - there are no one-eyed major league players. Of course, hitting a fastball is a peak-performance skill that very few individuals can master. To make the best use of two eyes, the bilateral symmetry of eyes and brain must be overridden through the use of the corpus callosum. This enables visual coordination between the eyes. Many animals are fully functional without the need to break the symmetry. To this end it would undoubtedly be instructive to study the structure of the corpus callosum in species that use these various strategies.

Resource Utilization

Brain Size

Neural Organization - continuum from direct connections and responses to matching of learned or remembered stimuli to an appropriate response

Energy Cost

Operating Modes - Hibernating, Sleep, Resting, Active

Recognizing Dead-End Technologies

Natural Selection has an uncanny way of trying and rejecting technologies that do not yield long-term benefits.

As an example, I contend that the fact that the wheel-and-axle does not occur in nature means that the approach is actually a dead-end. The fact that signals, nutrients and waste cannot pass through a fully rotating joint means that such a thing cannot have long-term utility for an organism.

When humans discovered and began using the wheel it was a wondrous, labor-saving device. However, it is likely that over the long term the overall energy costs and environmental pollution related to use of the wheel will make it non-viable. The inability to perform lubrication and maintenance across a fully-rotating joint means that we over-engineer the wheel to deal far beyond the worst-case usage. This means that non-degradable components outlast their normal application. The added burden, in the form of manufacturing costs and reduced resource availability, must be shared among all users.

Nature already has a stable of alternative options that do not involve these limitations. Sliding joints can provide load-bearing movement on two axes, are self-lubricating, self-repairing, capable of dynamic growth, allow unlimited passage of fluids and signals, and enable the static connection of musculature across the joint.

A corollary to this is the observation that nature will never use a rotary fan or impeller. If nature needs a pump it will either be based on peristalsis (the bowel) or the bellows for positive-displacement applications such as heart or lungs.

The natural choices exhibit incredibly high energy efficiency, are quiet, long-lasting, operate successfully at all scales as the organism grows, and are capable of functioning at speeds from zero up to their fluid-dynamic limit.

Natural components are created and assembled in-situ, on demand, at room temperature from a small menu of raw materials, adapted immediately to form and fit, and are, ultimately, easily degraded back into reusable materials. Conversely, every ball bearing ever created will end up in a landfill, taking their valuable, highly-refined raw materials with them.

These aspects should all be considered part of the “total cost of ownership” of a new technology. Nature has already factored in these line-items as part of the optimization between competing species. A wise designer would do well to emulate these successes.

In the case of Vision Systems, I contend that a similar dead-end technology is the shutter. No naturally-occurring vision system incorporates a shutter.

The invention of representational art constitutes a seminal development in human history. It allowed communication of new concepts across individuals and generations. The discovery of photography extended the concept in wondrous new directions by increasing the accuracy and reducing the energy expenditure involved in creating still images.

I believe that we are currently straining the applications far beyond their breaking-point by trying to maintain this frame-by-frame approach in two areas. First (as output), the use of movie-like sequences of still images to trick the eyes into perceiving an intended version of the real world. And second (as input), trying to interpret frame-by-frame input data from video cameras within machine vision systems.

Specifically we should evaluate the naturally-occurring vision systems and find better, faster and more resource-friendly methods of performing input and output of visual data.

No natural neural processing chain is synchronized. Dynamic events (sensations) propagate signals through neural pathways on demand. Parallel pathways propagate data at some inherent speed, independently, and only as needed.

Machine vision systems try to emulate this parallelism in various ways, given the limited available hardware in modern processors and communication channels. One example is image convolution - an attempt to perform the operations of parallel neurons.

Unfortunately, the approach is fatally flawed. The parallelism is an illusion, even recognizing the limitations of the "serial simulation of parallelism" imposed by the limited hardware. The entire process waits for the "next image". Then, like the starting gun at a horse race, all the "parallel" processing paths take their new data and try to deal with it. The vast majority of these processes will will have consumed time and energy but will yield no actionable results.

Richard Feynman would refer to this as *Cargo Cult Science*: all the trappings of science but none of the substance. Natural vision systems tightly couple their sensors to the associated neural processing, and this processing is is activated only on demand. Synthetic vision systems, on the other hand, provide merely the illusion of an equivalence. Ultimately, efforts to improve the illusion without addressing the fundamental flaw are counterproductive and doomed to fail.

All natural systems are inherently fault tolerant. Imperfections are dealt with at all stages of growth, training, and active use. This capability minimizes resource utilization and increases longevity.

Synthetic systems are, in general, anything but fault tolerant. Manufacturing processes strive for perfection and vast resources are devoted to refining these processes using such metrics as "process yield". For example, image sensors with a single defective pixel are discarded.

As another approach, memory arrays are typically designed with a limited number of alternative, redundant rows which can be enabled during manufacturing when a bad cell is detected in the primary array.

These are brute-force techniques which do not address the actual underlying problem, do not address defects encountered during the life of the product, and needlessly increase the overall cost of the system.

Natural systems tend to treat all defects as background noise with respect to the desired signal. The same noise-mitigation strategies that lead to improved processing of real, noisy, data from the environment will transparently deal with the noise introduced by defects within the organism.

Instead of requiring a perfect array of regularly spaced pixels, natural systems simply learn the relationship among the photoreceptors by observing the sequence of sensations encountered during normal vision. This on-going, dynamic updating of the understanding of the properties of the retina (for example) means that overall system performance is continually optimized over the life of the organism.

Furthermore, nature inherently allows for both the fovea and the blind spot. The retina has nothing even remotely similar to a regular array of equal-sized pixels. Therefore, it allows a high-density, high-resolution region to be combined with a wide field of view without developmental or computational penalty. And the blind spot introduced in the area of the interconnection between the movable sensor structure and the optic nerve does no material damage to the system performance, despite the resulting hole in the visual field.

The natural, continuous-learning process stands in sharp contrast to the modern synthetic vision techniques which tend to incorporate a “train it once at the factory, then use it forever in the field” paradigm.

Again, the current direction appears to represent an obvious dead-end.

Supporting Structures

Turning or Bobbing Neck or Body

Orbital Muscles

Eyelids and Blinking

Eyelashes, eyebrows and whiskers

Tears for moisture and cleaning

Visual Cortex - object recognition and memory

Motion Tracking

Ears and targeting audible positions - spatial awareness

> Interesting things sometimes make noises, and
> noises sometimes come from interesting things.

Learning to coordinate this experience is a survival trait.

Auditory cortex resolves amplitude, frequency and phase information. Some mammals are able to significantly alter the shape and orientation of the pinna to help resolve directionality. Environmental reflections as well as individual pinna reflections. Sound shadowing by the head. Human ear spacing and ~2 KHz threshold for phase vs. amplitude discrimination.

Echolocation. Bats and other mammals.

Understanding speech is always much easier if you watch the speaker. Synchronizing facial actions and gestures with prosody reduces guesswork and increases accuracy. Communication involves expressions as well as sounds. Just as audio feedback is critical in properly shaping speech with the desired properties when speaking, anticipating the mannerisms of another person forms the basis of a kind of “predictive keyboard” with likely choices of words or phrases pre-selected. This minimizes the computational load for routine speech and leaves the high-powered processing for nuance or unique situations. Examples would be recognizing irony, where the gestalt meaning may be opposite to the explicit meaning of any particular word or gesture. Nuance, emphasis and detail may come, in coordination, through any of the senses.

Phylogeny

PLANTS
consider phototropism

ALGAE
eyespots or stigma in photosynthetic organisms

PROTOZOA
are there single-cell examples of photosensitivity - seek/avoid light or changes in light

INSECTS
Compound Eyes
Optical Surface Cleaning Technique
Cockroaches and the dark
Bees and flowers
Praying Mantis and eye-stalks
Flies landing on ceilings
Moths to flames
Mayflies avoiding predators
Caterpillars
Moths

FISH
Underwater adaptations
Schooling
Maintaining observation of food/threat
Sharks and the lateral line
Pelagic creatures and bioluminescence

Mollusks
Octopi

AMPHIBIANS
Tadpoles
Frogs

REPTILES

Crocodilian Eye Placement
Water / Air Adaptation
Snakes in trees and deserts
Lizards
Diet - Insects vs. mammals
Infrared vision - pit vipers

BIRDS
Owls vs. sparrows
Male vs. Female

MAMMALS
Cats vs. Mice
Wolves vs. Antelopes
Hyenas
Whales
Humans

Summary

The adversarial evolution of species within a particular environmental niche will yield solutions to common problems that minimize costs in terms of energy and resource usage. Studying these evolved systems will allow us to access (1) a clearly bounded description of an important problem as well as (2) a candidate solution that has been well-vetted for viability. We may choose to implement our own variations to any particular natural solution. Nature provides a reference design, with established performance, for comparison.

Recognizing the existence of diverse solutions within apparently similar environmental niches may lead to unexpected variations in performance goals that may allow us to simplify our own designs. The presence of multiple solutions may also indicate alternative minima in the energy / resource utilization optimization formulae.

As new technological discoveries or inventions take place, it is often helpful to place them into a larger context. This Taxonomy may be used to reveal unforeseen applications and limitations related to similar technologies.

As an engineer, I cannot solve a problem that I do not know exists. If I have never experienced the problem, I do not understand the issues. However, if I am presented with a *solution*, I can often infer the problem. And given this information I can offer my own solutions.

Glossary of Technical Terms

Some terms are used in a technical sense that may be familiar to practitioners in a particular discipline. They may be either unfamiliar or used in a different sense by others.

adaptation - incremental evolutionary changes that make a particular structure or skill more suited to a particular environment

clade - Individuals descended from a common ancestor

clan - Group of interrelated families

corpus callosum - Wide nerve bundle that connects the left and right hemispheres of the brain of placental mammals

experience - the combination of routine stimuli and observed responses encountered by an individual and the resulting changes to a neural net. This makes the individual uniquely suited to respond to recent events and is the key to learned adaptation. Compare **training**.

familiarity - the ability of a neural network to build on **experience** and recognize subtle changes to frequently encountered objects or environments.

fovea - small portion of the retina with the highest acuity. About 1-2 degrees in visual field diameter in humans

learning - the process of using feedback and reinforcement to adjust the responses of a neural net

macula - portion of the retina responsible for the central portion of the visual field, as opposed to peripheral vision outside the macula

microsaccade - small jerk-like involuntary eye movements of 2 to 120 arc minutes, typically occurring during periods of visual fixation lasting several seconds

neural net - system of nerves that takes multiple stimuli and generates an output response based on the relative importance of each of the individual stimuli. The process of assigning those relative importances is referred to as **learning**. Stacking multiple neural nets sequentially such that responses of one net act as stimuli for the the next creates a **deep neural net**.

ontogeny - the development of an organism from fertilization to adulthood

optic disc - Portion of the retina where the optic nerve connects. It has no photosensors, and is referred to as the “blind spot”.

pareidolia - the tendency to perceive an object, pattern or meaning where there is none. Examples include seeing faces in objects or hearing voices in noise.

photosynthesis - process used by plants, algae and cyanobacteria to convert light energy, carbon dioxide and water into carbohydrates such as sugars and starches

phototropism - growth of an organism in response to light stimulus

pinna - external ear in mammals

saccade - quick, simultaneous movement of both eyes in the same direction. In humans rates of 700°/sec over a visual angle of 25°.

tapeta lucida - reflective membrane behind the retina of nocturnal mammals.

training - an explicit period of controlled stimuli and immediate feedback used to create standardized learned responses in a neural net. This training can be applied to a cohort of individuals and is expected to instill standardized responses. Compare **experience**.

vestibular system - sensory system in vertebrates that detects rotations and translations. Contributes to the sense of balance and spatial orientation

vestibulo-ocular reflex - use of the **vestibular system** to stabilize the field of vision independent of any visual stimulus. Triggers tracking of the visual field to counter motion of the head. May cause rotation of the eyeballs around the line of sight.

Appendix A - Species Data Questionnaire

Use this form to standardize information and notes related to natural vision systems. The outline of topics presented in this Taxonomy should provide interesting areas of exploration when preparing a Questionnaire. Conversely, discoveries made during the consideration of a particular species should inform revisions to the Taxonomy.

Species: **Developmental Stage:**

Describe the path of light through all structures until it reaches a photoreceptor.

Describe the photoreceptor location and sensitivity.

Describe the neural response network activated by this photoreceptor and group of receptors.

Describe the response mechanism triggered by this neural pathway.

Describe any feedback mechanisms related to this visual process.

Describe any specific support requirements of this visual process (i.e. specific nutrients).

Describe objective primary and secondary purposes of vision in this organism.

Describe the environment(s) in which this species operates.

Provide additional notes concerning novel structures of response situations or techniques.

