Ph.D. in Cognitive Science from the University of California at San Diego
Defense of the thesis *Active Perception* given on November 29, 2010.

# Précis of *Active Perception* by Nicholas John Butko, Ph.D.

## Overview and theoretical framework:

The aim of *Active Perception* is to contribute to the understanding and synthesis of perceptual intelligence. Attempts to synthesize intelligent skills in computational systems have typically succeeded in inverse proportion to the perceived difficulty of those skills. For example, winning complex games is difficult for humans but relatively easy for computers; meanwhile, "seeing" is an effortless endeavor for humans but is challenging for computers. This disparity indicates that there is still much to be understood about the fundamental nature of human intelligence, particularly perceptive intelligence.

Perceptive intelligence is an active process at many levels: neurons grow synapses and modulate their firing properties; attention sweeps its spotlight from place to place; eyes move; infants explore the effects that they cause; scientists design more informative experiments to understand the world; groups of individuals gather and coordinate information to form a shared understanding. Although there are unique peculiarities at each of these levels, there are also basic commonalities.

*Active Perception* frames perception as a game that we play with our environment. In games, we choose the moves that maximize our expected chance of winning. In perception, we act to best understand the world.

To formalize perceptive intelligence as a game, the following questions must be answered: (1) What is the goal of the game? How do we measure how much we "understand the world"? (2) How complex is the game, and how can we deal with that complexity? (3) How do we learn to play the game, when nobody gives us the rules? (4) Why is it beneficial to "understand the world", anyway? *Active Perception* proposes answers to each of these questions.

(1) *What is the goal of the game? How do we measure how much we "understand the world"?* Understanding can be measured using mutual information, which measures the reduction in uncertainty caused by a sequence of actions. Actions are considered good for perception if they optimally reduce uncertainty, *i.e.* lead to maximum information gain.

(2) *How complex is the game, and how can we deal with that complexity?* The complexity of a game can be measured as its branching factor. This is a product of the moves, or actions, that a player can make and the number of possible outcomes of those actions. For example, in a game like chess there are, on average, 30 moves available to each player. Thus, the branching factor is about 30 actions x 30 responses, or 900 outcomes. In perceptual domains, the

space of actions is typically low dimensional and continuous (for example, you can move your eyes left-right and up-down), and the space of outcomes is high dimensional and continuous (every retinal ganglion cell gives some response). In Chapter 5 of *Active Perception*, we consider a discretized retina with about 400 pixels. If these pixels had binary responses, there would be $2^{400}$ possible sensory outcomes for each possible eye movement. Because the model retina operates in a continuous sensor space, the branching factor is in fact even higher.

In chess as well as in other classical AI games, it is difficult, but possible, for a computer to look many moves ahead and evaluate which decision is expected to maximize the chance of winning. However, in some perceptual domains, it is simply impossible for any computer to look even a single move into the future. Yet, we demonstrate approximate algorithms that learn to solve this problem, which achieve identical performance to the explicit planner in simple toy problems and provide useful solutions to problems for which explicit planners fail. The key is to employ modern reinforcement learning algorithms that are driven by information gain as an intrinsic reinforcer. These algorithms succeed in perceptual domains where explicit planners fail, and they advance the state of the art in machine perception.

(3) *How do we learn to play the game, when nobody gives us the rules?* The rules of a game can be encoded as transition probabilities in a Markov decision process (MDP). Robots and humans are embedded in a sensory environment governed by physical laws and contain a rich statistical structure. By observing the stochastic relationship between action and effect, a robot can learn to make sense of its own sensorimotor experiences. This kind of learning is the focus of Chapters 6 and 7 of *Active Perception*. For example, we show how a robot can learn the statistical relationship between signals sent to eye motors and the resulting translation of pixels across its retina. This learning is grounded in low level sensor data and requires no external teacher. We show how this self-taught knowledge can be used to optimally coordinate multiple gaze components, such as the head and eyes, in order to fixate visual targets with maximum accuracy. Notably, the mathematically optimal learning rules and control laws share many features with pathways in primate neural circuitry.

(4) *Why is it beneficial to "understand the world", anyway?* Historically, the goals of organisms have been understood in terms of concrete reinforcers, such as food. In this context, reducing uncertainty may seem to be a strange motivator. In contrast, information is a cognitive reward. Recent results in neurophysiology suggest that neural reward systems give explicit positive feedback for behaviors that lead to a reduction in uncertainty. In *Active Perception*, we focus on the concrete benefits of information as a driver for different kinds of learning. In Chapters 2 and 3, we show that the information contained in the statistics of visual data can be used to learn useful visual representations and can be used to efficiently allocate attention. In Chapters 4 and 5, we show that information gain can reinforce behaviors that maximize perceptual efficiency. In Chapter 1, we present a coherent theoretical framework to discuss different ways that information can be used to drive perceptual learning.

# Brief summary of studies:

*Active Perception* contains six related research studies that span multiple levels of cognitive science. This breadth is possible because the underlying computational framework focuses on the commonalities of all levels. Studies 1 to 6 correspond to Chapters 2 to 7 of *Active Perception,* respectively, while Chapter 1 contributes a unifying theoretical framework.

Study (1) considers a population of active neurons with information theoretic objectives. They modify internal and synaptic parameters to maximize their capacity as information channels. We show that these model neurons exhibit firing rate patterns that have been previously reported in empirical studies as well as receptive field properties that have been demonstrated in purely computational approaches. Our model neurons solve a non-linear independent components problem and learn Gabor-filter-like receptive fields. The neural adaptation mechanism is simple and local in time and space, which are necessary for a plausible biological mechanism. This work suggests new algorithms for unsupervised visual feature extraction, a basic building block of visual object recognition systems.

Study (2) illustrates how a computational system can estimate, in real time, low-level statistics of the visual environment and the amount of information present, which has previously been used to model visual salience. The real time implementation leverages efficient spatio-temporal filtering techniques and efficient inference techniques. This enables implementation and deployment of a real-time model of attention, thereby facilitating an empirical evaluation of the usefulness of visual salience. In a field study, we tested this model in a robot at an early childhood education center, and we found that the salience model directs the robot's attention to regions of the scene that contain people without requiring any specific knowledge of the appearance of humans. Although it is meant to model human visual attention, our system shares many algorithmic similarities to the popular SIFT key point detection algorithm, while operating at a rate that is about eighteen times faster. In the future, we would like to compare the stability of the key points that are extracted from these two methods.

Study (3) investigates contingency detection, a process known to be fundamental in the development of social interaction in infants. We model the process of infants learning to learn. We conceptualize this as the process of learning to gather information in an efficient manner. We show that this framework accounts for observed infant developmental trajectories. We consider an agent endowed with a single binary sensor and a single binary actuator. The agent uses these sensorimotor capabilities to optimally probe its environment and learn whether a second social agent is present as quickly as possible. We fit parameters of this model empirically by studying humans' social interactions with robots. Given the empirical statistics of social interaction, the optimal decision rule can be computed analytically. This allows us to compare the behavior of ten-month-old infants to the optimal controller. The performance of human infants and the optimal controller are virtually identical in many important aspects, thereby raising the question of how infants learn such good learning strategies. We show that information gain, or reduction of uncertainty, can be used as a reinforcing signal to drive learning. Such learning could account for developmental shifts that are observed in infant studies. At the time of this study, it was unknown if information gain was a biologically viable reinforcing signal. Since this study, it was demonstrated in rhesus macaque monkeys that dopamine based reward systems actively reward behaviors that lead to information gain, regardless of whether the gained information indicates that the monkey will receive an external reward.

Study (4) focuses on the problem of visual search, which we analyze using the framework of stochastic optimal control, driven by information reward. In practice, this task is difficult due to the huge branching factor. Computing the expected information gain for even the next single fixation is computationally prohibitive. Instead, we show that a modern reinforcement learning algorithm can use information as a reward signal to optimize parameters of a neural network. In simulated visual search tasks, this network learns to outperform controllers that are optimal for a one-step lookahead; yet it incurs a fraction of the computational cost. We then construct a digital retina based on this control policy. The retina is able to detect faces in images faster than current, state of the art computer vision algorithms. Thus, a model based on human physiology and optimal control yields a better AI program. While it was previously assumed that the optimal search strategy is to fixate the region of space most likely to contain the target, we show that this strategy can be suboptimal. Our model predicts changes in the optimal search strategy depending on the visual properties of the search target, *e.g.* level of contrast, as well as based on its movement dynamics. Our model, therefore, suggests hypotheses for future psychophysical studies. Our model also predicts changes in the optimal search strategy depending on the characteristics of the eye, *e.g.* foveated or not. Thus, robots with non-foveated cameras may benefit from different eye-movement properties than humans. Future studies could address whether human observers prefer robots that move their eyes in the same manner as humans, thus appearing intelligent, or move their eyes in the manner that is optimal for the robot, thus exhibiting intelligence.

Study (5) proposes a method in which a robot learns how to coordinate its eyes and head as it learns to search for objects. This entails learning about the physical parameters of the robot's oculomotor system. We show that learning to look at visual targets contains a rich problem structure that relates sensory experience, motor experience, and development. By encoding this intrinsic problem structure in a generative model, we show how an optimal observer should arbitrate different sources of uncertainty to discover how its sensors and actuators are related. We implement our approach on three different robots, and we show that all of them can quickly learn reliable, intentional looking behavior without access to any information beyond their own experiences. The learning is robust to differences in robot morphology, damage to robots such as missing limbs, and extreme perturbations such as kidnapping. Finally, we propose a principle by which multiple gaze components, *e.g.* head and eyes, should be combined based on a principle of maximum accuracy in the presence of signal dependent noise. The solution to the optimization problem accounts for undershooting, which is reliably observed in human saccades when the head is fixed and only the eyes are allowed to move; it further makes predictions that such undershooting should be mitigated when the head is free to move and thus multiple gaze components are allowed to coordinate. The optimal gaze coordination rule in this study is for a single saccade. In the future, we would like to use reinforcement learning to coordinate the motion of the head and eyes in the context of multiple saccades, where no closed form solution is known to exist.

Study (6) demonstrates that joint statistics of acoustic and visual modalities can be used to teach an infant to recognize visual objects, such as people. We deployed the learned acoustic social contingency detection behavior from study (3) in a physical robot with the appearance of a human infant. This baby robot actively vocalizes to probe its environment for social

interlocutors. We then invited subjects to interact with the robot. When the robot detected acoustic contingencies, it saved images and marked them as “social interaction present.” When it detected no contingencies, it saved images and marked them as “social interaction absent.” Then, the robot applied a previously published visual learning technique, called segmental Boltzmann fields (SBFs). SBFs are a learning algorithm designed to work on weakly supervised problems, *i.e.*, problems in which we are told whether or not the category of interest is rendered in the image, but we are not told where it is rendered. Thus, the goal is to learn to localize objects from the target class in novel images. With less than six minutes of experience, sampled from ninety minutes of interaction with the world, the robot learned to find people in novel images. In addition, the baby robot developed a preference for drawings of human faces over drawings of non-faces, even though it had never been exposed to such schematic face drawings before. During the six minutes of training, the robot was never told whether people were present in the images or whether people were of any particular relevance at all. It simply discovered that to make sense of the images and sounds it received, it was a good idea to use feature detectors that happened to discriminate the presence of people. The results illustrate that visual preferences that are typically investigated in human neonates can be acquired very quickly, in a matter of minutes. Previous studies that were thought to provide evidence for innate cognitive modules may actually be evidence for rapid learning mechanisms in a neonate brain that is exquisitely tuned to detect the statistical structure of the world. While this study used contingency cues to indicate the presence of caregivers, there are many potential such cues, such as motion, salience, or touch. It remains an open question which of these other cues could also suffice to bootstrap visual learning without a teacher.

# Findings, and contributions to cognitive science:

**Chapter 7:**

- We show that a baby robot can learn about the visual appearance of humans from six minutes of data about acoustic contingency cues.
- The robot exhibits the same preference to schematic face and non-face stimuli as forty-minute-old neonates.
- The robot develops the same preference for people in its environment over strangers as one-day-old neonates.

Morton & Johnson argued that infants may be born with an innate knowledge of what their species looks like, and they expressed skepticism that the preferences observed in neonates to face-like sketches could be learned from the small amount of available data. We don’t claim that neonates are learning using the same algorithm as our robot, but we can make a strong claim that there is enough visual information in the infant's environment to support rapid visual learning.

**Chapter 6:**

- We derive optimal inference rules for a robot to learn how to look at visual targets.
- The learning is grounded in raw sensorimotor data and requires no external teacher.
- The optimal inference rule entails terms directly analogous to corollary discharge, and it also entails predicting the specific sensory outcome of a saccade, which has previously been observed in monkey lateral intraparietal cortex. (Study 5).
- We derive an optimal fixation rule for coordinating the heads and eyes to fixate visual targets with maximum accuracy.
- The optimal fixation rule leads to an undershoot bias, which is also observed in humans when the head is fixed.
- The optimal fixation rule mitigates undershooting when the head and eyes are allowed to coordinate, which becomes a hypothesis for future human experiments.

**Chapter 5:**

- We build a digital eye that scans images to detect faces twice as fast as previous methods.
- The digital eye is based on psychophysical models of human active search using foveated vision.
- The digital eye achieves high performance by choosing search targets using a simple neural network trained with reinforcement learning and an information based reward signal.
- The neural network chooses the next fixation target much faster than previously proposed methods, and it exhibits better performance.
- Our method overcomes the huge branching factor of perceptual control domains.
- We show that optimal visual search varies as a function of sensor characteristics and target dynamics. For example, the optimal search strategy changes for human eyes *vs.* a robot's camera and for static targets *vs.* moving targets.
- Previous investigations have assumed that the optimal strategy for visual search is to look at the locations with the highest probability of containing the search target; we show that this strategy is suboptimal.
- We show that a behavior that was previously interpreted as "forgetting" in human subjects can be explained as optimal inference when the world is likely to change in uncertain ways, *e.g.* visual targets can move even when you're not looking at them.

**Chapter 4:**

- We build a contingency detection system that actively discovers acoustic social interactions in real time in real life environments, even very noisy ones, by maximizing information gain.
- We demonstrate that ten-month-old human infants' behavior in social contingency experiments is consistent with the behavior we'd expect if their goal was to maximize the information they receive.

- The optimal controller exhibits the same turn-taking behaviors that are observed in ten-month-old infants.
- The optimal controller exhibits the same timing signatures as behaviors observed in ten-month-old infants.
- We show that information gain is a sufficient reward signal to drive the learning of optimal probing behaviors in a ten-month developmental time frame.
- We predict that the human neural reward systems are sensitive to uncertainty reduction.
- After our initial study was published, it was shown that midbrain dopamine systems thought to be responsible for reward-based learning respond strongly to the reduction in uncertainty.

**Chapter 3:**

- We implement the first published real-time salience algorithm.
- We deploy this algorithm in real life conditions and show that low-level salience biases a robot camera toward relevant aspects of daily life, *e.g.* humans.
- The algorithm is based on a model that also accounts for human search asymmetries that are observed in psychophysical experiments.
- Our approach matches the performance of models that are fit to human data on the task of predicting where humans will fixate.

**Chapter 2:**

- We demonstrate a novel way to learn sensory transformations that are similar to those observed performed by primary visual cortex.
- The learning algorithm uses computations that are more straightforward to implement in biological circuits than previous information theoretic approaches.
- Our approach uses model neurons that follow an information gradient to maximize their capacity as information channels.

**Chapter 1:**

- We review previous approaches to information maximization in various fields of cognitive science including primary sensing systems, attention systems, and information foraging behavior.
- We present and compare these disparate views in a unified framework.
- We present central challenges in active perception.