Sensor Fusion in Distributed Cortical Circuits

Mohsen Firouzi

Graduate School of
Systemic Neurosciences
LMU Munich

Dissertation at the
Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universität München

September, 2019

Supervisors:

Prof. Dr. Jörg Conradt,
KTH Royal Institute of Technology,
Division of Computational Science,
Stockholm, Sweden

Prof. Dr. Stefan Glasauer,
Ludwig-Maximilians-University of Munich
Graduate School of Systemic Neurosciences,
Munich, Germany

Examiners:

Prof. Dr. Bernhard Wolfrum,
Technical University of Munich
Neuroelectronics Group,
Munich, Germany

Prof. Dr. Zhuanghua Shi,
Ludwig-Maximilians-University of Munich
Department of Psycholog,
Munich, Germany

External Reviewer:

Prof. Dr. Jeffrey L. Krichmar,
University of California, Irvine,
Cognitive Anteater Robotics Lab,
Irvine, USA

Date of Submission: 15/10/2019
Date of Defense: 27/03/2020

To the most precious being in my life, to my Mum.

Preface & Acknowledgement

I did my B.Sc. and M.Sc. in Electrical Engineering at Sharif University of Technology. At the first month of my postgraduate study, I was the first student at the whole Department who already knew at which lab I want to work: *Artificial-Creatures-Lab, Sharif University of Technology*. The head of this lab was a great disciplined professor in AI and Cognitive Science, the one whom I wanted to work with. I was so excited to know more and more about theories in Computational Intelligence and Brain day by day. Understanding the brain had become an inerasable part of my passion. Even though I had been working in a quite different discipline after graduation – satellite technology, I still spent sometimes with my team-leader discussing about building an artificial brain in silicon. In February 2012, I read a flyer about a PhD program in Systemic Neuroscience, Ludwig-Maximilian-University of Munich. I already knew it is not going to be an easy life. Not as I used to already have, a good job, good friends, happy life and a wonderful family. A friend always used to quote me "*A PhD is going to be an enduring daring adventure*". But, I was decided to pursue my PhD in this Department, and to open up a new Chapter of my life somewhere in Munich. Here my journey began. I should admit that after years of being in this road, I would never forget any single moments of that. The cheers, the grief; the anger, the smiles; the long nights spending in the Lab and sleeping in a couch, beautiful snowy nights of *Geschwister-Scholl-Platz*, and wonderful springs of *Biozentrum*; the moments of failure and success, the long days of fighting with a sickness and disappointment. I have learnt many things beyond what I imagined. Something about myself, about life, and about people. And it is something beyond a simple academic life. Now I have a new family: GSN^{LMU}, and I feel so proud of it.

I could not definitely finish this journey without the great support of many people. Hereby, I would like to thank *Dr. Jörg Conradt* and *Dr. Stefan Glasauer* for supervising this project and their valuable comments. It is a great privilege for me to express my warmest gratitude to *Graduate School of Systemic Neurosciences*, particularly Mrs. *Lena Bittl* and Mrs. *Stefanie Bosse* for all their supports and encouragements, without which the fulfillment of my doctorate would have never been possible. It is indeed an honor to be part of such a wonderful family. GSN deserves credits not just for its scientific contributions and reputation, but also for such a family-like attitude. For a young international student, there is nothing better than being amongst such people. Thank you GSN, for being the reason of writing these lines tonight. I really do not know how to phrase my appreciations to my family particularly to my *Mum* and my beloved Sister *Zohre*, for what they have

done for me during these years. Thanks for being always there to grab ahold of me in hard moments and to reassure me towards the end. I appreciate the time my lovely friend *Dr. Hesam Sagha* spent to proofread my dissertation, the wonderful teamwork with *Dr. Christoph Richter*, and helpful discussions with *Dr Mathew Cook* and *Prof Alexander Pouget*. At the end, I am cordially thankful to *Bernstein Center for Computational Neuroscience – Munich*, for funding this project, deploying a wonderful scientific atmosphere for my academic career, and for allowing me to be part of this society as a junior scientist.

Mohsen Firouzi,
5th September 2019, Munich

Table of Content:

Abstract

The substantial motion of the nature is to balance, to survive, and to reach perfection. The evolution in biological systems is a key signature of this quintessence. Survival cannot be achieved without understanding the surrounding world. How can a fruit fly live without searching for food, and thereby with no form of perception that guides the behavior? The nervous system of fruit fly with hundred thousand of neurons can perform very complicated tasks that are beyond the power of an advanced supercomputer. Recently developed computing machines are made by billions of transistors and they are remarkably fast in precise calculations. But these machines are unable to perform a single task that an insect is able to do by means of thousands of neurons. The complexity of information processing and data compression in a single biological neuron and neural circuits are not comparable with that of developed today in transistors and integrated circuits. On the other hand, the style of information processing in neural systems is also very different from that of employed by microprocessors which is mostly centralized. Almost all cognitive functions are generated by a combined effort of multiple brain areas. In mammals, Cortical regions are organized hierarchically, and they are reciprocally interconnected, exchanging the information from multiple senses. This hierarchy in circuit level, also preserves the sensory world within different levels of complexity and within the scope of multiple modalities. The main behavioral advantage of that is to understand the real-world through multiple sensory systems, and thereby to provide a robust and coherent form of perception. When the quality of a sensory signal drops, the brain can alternatively employ other information pathways to handle cognitive tasks, or even to calibrate the error-prone sensory node. Mammalian brain also takes a good advantage of multimodal processing in learning and development; where one sensory system helps another sensory modality to develop. Multisensory integration is considered as one of the main factors that generates consciousness in human. Although, we still do not know where exactly the information is consolidated into a single percept, and what is the underpinning neural mechanism of this process?

One straightforward hypothesis suggests that the uni-sensory signals are pooled in a ploy-sensory convergence zone, which creates a unified form of perception. But it is hard to believe that there is just one single dedicated region that realizes this functionality. Using a set of realistic neuro-computational principles, I have explored theoretically how multisensory integration can be performed within a distributed hierarchical circuit. I argued that the interaction of cortical populations can be interpreted as a specific form of relation satisfaction in which the information preserved in one neural ensemble must agree with incoming signals from connected populations according to a relation function.

This relation function can be seen as a coherency function which is implicitly learnt through synaptic strength.

Apart from the fact that the real world is composed of multisensory attributes, the sensory signals are subject to uncertainty. This requires a cortical mechanism to incorporate the statistical parameters of the sensory world in neural circuits and to deal with the issue of inaccuracy in perception. I argued in this thesis how the intrinsic stochasticity of neural activity enables a systematic mechanism to encode probabilistic quantities within neural circuits, e.g. reliability, prior probability. The systematic benefit of neural stochasticity is well paraphrased by the problem of Duns Scotus paradox: imagine a donkey with a deterministic brain that is exposed to two identical food rewards. This may make the animal suffer and die starving because of indecision. In this thesis, I have introduced an optimal encoding framework that can describe the probability function of a Gaussian-like random variable in a pool of Poisson neurons. Thereafter a distributed neural model is proposed that can optimally combine conditional probabilities over sensory signals, in order to compute Bayesian Multisensory Causal Inference. This process is known as a complex multisensory function in the cortex. Recently it is found that this process is performed within a distributed hierarchy in sensory cortex. Our work is amongst the first successful attempts that put a mechanistic spotlight on understanding the underlying neural mechanism of Multisensory Causal Perception in the brain, and in general the theory of decentralized multisensory integration in sensory cortex.

Engineering information processing concepts in the brain and developing new computing technologies have been recently growing. Neuromorphic Engineering is a new branch that undertakes this mission. In a dedicated part of this thesis, I have proposed a Neuromorphic algorithm for event-based stereoscopic fusion. This algorithm is anchored in the idea of cooperative computing that dictates the defined epipolar and temporal constraints of the stereoscopic setup, to the neural dynamics. The performance of this algorithm is tested using a pair of silicon retinas.

Zusammenfassung

Die wesentliche Bewegung der Natur besteht darin, auszubalancieren, zu überleben und Perfektion zu erreichen. Die Evolution in biologischen Systemen ist eine wesentliche Signatur dieser Quintessenz. Überleben kann nicht erreicht werden, ohne die umgebende Welt zu verstehen. Wie kann eine Fruchtfliege leben, ohne nach Nahrung zu suchen, und damit ohne eine Form der Wahrnehmung, die das Verhalten steuert? Das Nervensystem der Fruchtfliege mit hunderttausenden von Neuronen kann sehr komplizierte Aufgaben erfüllen, die die Möglichkeiten eines modernen Supercomputers übersteigen. Neu entwickelte Rechenmaschinen bestehen aus Milliarden von Transistoren und sind bei präzisen Berechnungen bemerkenswert schnell. Aber diese Maschinen sind nicht in der Lage, eine einzige Aufgabe zu erfüllen, die ein Insekt mit Hilfe von Tausenden von Neuronen erledigen kann. Die Komplexität der Informationsverarbeitung und Datenkompression in einem einzigen biologischen Neuron und neuronalen Schaltkreisen ist nicht vergleichbar mit der, die heute in Transistoren und integrierten Schaltkreisen entwickelt wird. Andererseits unterscheidet sich die Art der Informationsverarbeitung in neuronalen Systemen auch sehr von der Art der Informationsverarbeitung in Mikroprozessoren, die meist zentralisiert ist. Fast alle kognitiven Funktionen werden durch die kombinierte Anstrengung mehrerer Hirnareale erzeugt. Bei Säugetieren sind die kortikalen Regionen hierarchisch organisiert, und sie sind wechselseitig miteinander verbunden und tauschen die Informationen von mehreren Sinnen aus. Diese Hierarchie auf der Ebene der Schaltkreise bewahrt auch die Sinneswelt innerhalb verschiedener Komplexitätsebenen und im Rahmen mehrerer Modalitäten. Der wichtigste Verhaltensvorteil besteht darin, die reale Welt durch mehrere Sinnessysteme zu verstehen und dadurch eine robuste und kohärente Form der Wahrnehmung zu ermöglichen. Wenn die Qualität eines sensorischen Signals abnimmt, kann das Gehirn alternativ andere Informationswege nutzen, um kognitive Aufgaben zu bewältigen oder sogar den fehleranfälligen sensorischen Knoten zu kalibrieren. Das Säugetiergehirn nutzt auch einen guten Vorteil der multimodalen Verarbeitung beim Lernen und bei der Entwicklung, wobei ein sensorisches System die Entwicklung einer anderen sensorischen Modalität unterstützt. Die multisensorische Integration wird als einer der Hauptfaktoren betrachtet, der beim Menschen Bewusstsein erzeugt. Obwohl wir noch immer nicht wissen, wo genau die Informationen zu einer einzigen Wahrnehmung zusammengeführt werden, und was der zugrunde liegende neuronale Mechanismus dieses Prozesses ist.

Eine einfache Hypothese besagt, dass die uni-sensorischen Signale in einer ploy-sensorischen Konvergenzzone gebündelt sind, was eine einheitliche Form der Wahrnehmung schafft. Aber es ist schwer zu glauben, dass es nur eine einzige dedizierte Region gibt, die diese Funktionalität verwirklicht. Mit Hilfe einer Reihe realistischer

neuroinformatischer Prinzipien habe ich theoretisch untersucht, wie eine multisensorische Integration innerhalb eines verteilten hierarchischen Schaltkreises durchgeführt werden kann. Ich argumentierte, dass die Interaktion kortikaler Populationen als eine spezifische Form der Beziehungszufriedenheit interpretiert werden kann, bei der die in einem neuronalen Ensemble erhaltene Information mit eingehenden Signalen von verbundenen Populationen gemäß einer Beziehungsfunktion übereinstimmen muss. Diese Beziehungsfunktion kann als eine Kohärenzfunktion angesehen werden, die implizit durch synaptische Stärke gelernt wird.

Abgesehen von der Tatsache, dass die reale Welt aus multisensorischen Attributen besteht, sind die sensorischen Signale mit Unsicherheit behaftet. Dies erfordert einen kortikalen Mechanismus, um die statistischen Parameter der sensorischen Welt in die neuronalen Schaltkreise einzubeziehen und die Frage der Ungenauigkeit der Wahrnehmung zu behandeln. Ich habe in dieser Arbeit argumentiert, wie die intrinsische Stochastizität der neuronalen Aktivität einen systematischen Mechanismus zur Kodierung probabilistischer Größen in neuronalen Schaltkreisen ermöglicht, z.B. Zuverlässigkeit, Vorwahrscheinlichkeit. Der systematische Nutzen der neuronalen Stochastizität wird gut durch das Problem des Duns-Skotus-Paradoxons umschrieben: Stellen Sie sich einen Esel mit einem deterministischen Gehirn vor, der zwei identischen Futterbelohnungen ausgesetzt ist. Dies kann dazu führen, dass das Tier aufgrund von Unentschlossenheit leidet und verhungert. In dieser Arbeit habe ich ein optimales Kodierungsgerüst eingeführt, das die Wahrscheinlichkeitsfunktion einer Gauß-ähnlichen Zufallsvariablen in einem Pool von Poisson-Neuronen beschreiben kann. Danach wird ein verteiltes neuronales Modell vorgeschlagen, das bedingte Wahrscheinlichkeiten über sensorische Signale optimal kombinieren kann, um die Bayes'sche multisensorische kausale Inferenz zu berechnen. Dieser Prozess ist als komplexe multisensorische Funktion im Kortex bekannt. Kürzlich wurde festgestellt, dass dieser Prozess innerhalb einer verteilten Hierarchie im sensorischen Kortex durchgeführt wird. Unsere Arbeit gehört zu den ersten erfolgreichen Versuchen, die ein mechanistisches Rampenlicht auf das Verständnis des zugrunde liegenden neuronalen Mechanismus der multisensorischen kausalen Wahrnehmung im Gehirn und allgemein auf die Theorie der dezentralisierten multisensorischen Integration im sensorischen Kortex werfen.

In jüngster Zeit sind die Konzepte der technischen Informationsverarbeitung im Gehirn und die Entwicklung neuer Computertechnologien gewachsen. Neuromorphes Engineering ist ein neuer Zweig, der diese Aufgabe übernimmt. In einem speziellen Teil dieser Arbeit habe ich einen neuromorphen Algorithmus für die ereignisbasierte stereoskopische Fusion vorgeschlagen. Dieser Algorithmus ist in der Idee des kooperativen Rechnens verankert, das die definierten epipolaren und zeitlichen

Beschränkungen des stereoskopischen Aufbaus der neuronalen Dynamik vorgibt. Die Leistung dieses Algorithmus wird mit Hilfe eines Paares von Silikon-Netzhäuten getestet.

Chapter 1

Introduction

"Biology gives you a brain. Life turns it into a mind."
— Jeffrey Eugenides, (1960 -)

1.1 Theory of Brain and Computation

History of science has been always dealing with unknown phenomena and complicated dilemmas that endangered our survival, e.g. plagues, illness epidemics, or challenged our curiosity and ambition to live longer and to push the frontiers of our knowledge towards a brighter future. We used to make theories about unknowns at the first place we face with it. Ancient sailors and explorers made fiction stories about sea trolls living in far seas to demonstrate the difficulty of reaching deep oceans and sailing across Atlantic. For centuries people in Europe believed that the sun orbits around earth, making the day-night cycle; or earth is carried by giant elephants. Similarly, there have been many different theories about human intelligence. How it is emerged and where it comes from. Is it exclusively generated by a biological organ? What is the reason for mental diseases, and how can they be cured?

Despite many open questions about the human brain, today we know a tremendous amount of facts about it. But it was not the case over past generations, and it is not developed overnight. Very recent archeological discoveries in North Africa revealed that ancient humans performed skull trepanation over 7000 years ago, perhaps for medical purposes [Jórdeczka 2016]. Maybe the ancient doctors might have been investigating the reason of some diseases caused by brain deficits. However, it is believed that the trepanation could have been used also for religious and magical purposes [Jórdeczka 2016]. Even in the middle ages, some doctors believed that opening the skull would release the satanic beings that would infect the patient and cause madness (FIGURE 1-1, *the cutting stone* painting). The most notable ancient scholar who described the brain as a center of sensation and intelligence is Greek physician *Hippocrates* (*460-379 B.C.*). He argued that the anatomy should be correlated with the function; since the sensation organs like eyes, ears, nose and tongue are all located in our head and they send fibers

FIGURE 1-1

"*The Extraction of the Stone of Madness*" or "*Cure of Folly*" painted by *Hieronymus Bosch* (1488–1516), displayed in the *Museo Del Prado* in *Madrid*. Artistic depiction of medieval people false belief. It depicts the trepanation procedure in the Middle Ages and the painter is ridiculing the false knowledge of his doctor (the man wearing a funnel hat) [Foucault 2004].

into the skull, the brain should be the source of human sensation and feeling. He also added that what we see, hear, taste, and the knowledge we acquire are all emerged by an organ inside the skull, that we call *Brain* nowadays. However, this function for the brain was not accepted by all scholars at that time. *Aristotle* (*384-322 B.C.*), the famous Greek philosopher, thought that the heart, not the brain, is the center of feeling and wisdom. *Aristotle*'s theory was: "*the brain is nothing more than a supplementary organ for heart, cooling the blood circulation*". There are drawings left by one of the first prominent physicians *Galen* (*130-220*), an ancient roman physician, who adhered to the *Hippocrates* theory, studied sheep brain. He made a distinction between two main parts of sheep brain, *cerebrum* and *cerebellum*. Then, he stated that the cerebrum must be the receiver of sensations, and cerebellum should command muscles.

During the Dark Age in Europe, people believed that the madness is caused by a demonic creature. After *Galen's* reports, for more than thousand years, no significant

development or scientific experiments about the brain is reported in west. During this time, Islamic and Arab-Persian scientists took up the flame of scientific development for about ten centuries. The first dedicated psychiatric hospitals were built around each corner of Islamic world (*Baghdad* in 705, *Cairo* in 800, *Damascus* and *Aleppo* in 1270), indicating the need for understanding human mental health [Syed 1981]. *Ibn al-Haytham* (965-1040) was the first scientist to report that vision should be perceived in the brain rather than the eyes (in "*Book of Optics*"). He argued that personal experiences affect what people see, or in other words, visual perception is a subjective feeling that can be influenced in the brain [Steffens 2006]. This theory is in line with modern theories in visual perception. *Al-Biruni* (*973-1048*) was a pioneer in experimental *psychology*, as was the first who empirically explained the concept of *reaction time* (taken from one of his lectures, translated to English):

"*Not only is every sensation attended by a corresponding change localized in the sense-organ, which demands a certain time, but also, between the stimulation of the sense-organ and perception an interval of time must elapse, corresponding to the transmission of stimulus for some distance along the nerves.*"

Avicenna (*in Persian, Ibn-Sina; 980-1037*), the famous Persian physician and philosopher, discovered the *cerebellar vermis,* that he named *vermis,* and the *caudate nucleus*, that he named *tailed nucleus*, the terms which are still used in modern neurophysiology [Aydin 2001]. Moreover, he was the first scientist who specifically reported the cause of some intellectual dysfunctions as potential deficits in the *frontal lobe* (which mediates common sense and reasoning) [Theodore 2006].

During renaissance, a growing movement began in Europe to develop new techniques in biology, medicine, experimental physics, and mechanics. After inventing mechanical machines, *René Descartes* (*1596-1650*) advocated the theory of "*brain as a mechanical machine*". Resembling hydraulically controlled machines, he believed that the neural fibers carry fluid to communicate with limbs and muscles. However, he thought that this mechanism can just explain those behaviors that human shares with animals. Later at early 18th, this idea was replaced with an alternative theory: "*the brain is an electrical machine*", where neural fibers convey electricity

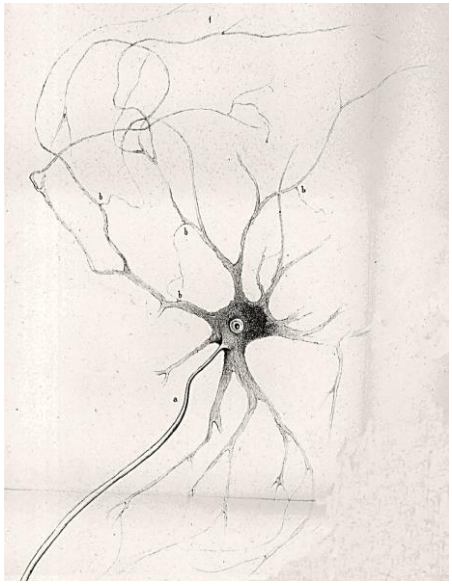


FIGURE 1-2

Drawing made by German anatomist, *Otto Deiters (*1834-1863*)*. It shows a single nerve cell and its *neurites* (*dendrites and axon*), taken from [Clarke and O'Malley 96].

rather than micro-fluid. At mid-19th with the advent of microscope, a breakthrough in understanding brain structure happened. At this time biologists could identify the *nerve cells* and *neurites* (FIGURE 1-2). Yet the revolutionary point needs to wait until *Cajal's* theory of *neuron doctrine,* where the nerve cells adhere to cell theory in biology. Before Cajal made his notion, neurites are thought to be like blood vessels and micro-channels that connect cells. In contrast with this theory, *Cajal* argued that nerve cells (*neurons*) are the elemental computational units of human brain which communicate using contacts (*synapses*) rather than a continuous reticulum. In other words, he stated that neurons are distinct cells, specialized to collect, to convey, to exchange, and to integrate *information*. Thus, to understand the brain, we need to understand the functions of neurons. *Cajal* is not the only scientist who contributed in developing *neuron doctrine*. However, this theory is coined by his name, and that well deserved the Nobel Prize of physiology and medicine in 1906.

Over the last century, thousands of brains are devoted to understand many facts about a single neuron, how morphology is correlated with functionality, how a neuron codes information and how the information is exchanged and stored in synapses, and what is the behavioral equivalence of the neural activities? But, complex behaviors are clearly not emerged by a single neuron. Human brain comprises several distinct parts namely cerebrum, cerebellum, thalamus, and brain stem; each part is composed of a complex layered structure of neurons. To study the mechanics of this complex machine, it is required to break it down into pieces, and to approach it from different levels of analysis. This mission today is handed over to multiple disciplines that are all appreciated to solve pieces of this puzzle, from molecular and cellular neuroscience to system and cognitive neuroscience.

The general scope of this thesis is *System neuroscience* that focuses on understanding the brain in circuit and system level. Brain can be divided into many subsystems with specialized circuitry and the style of information processing that generate particular functionalities e.g., vision, motor control, attention. When it comes to system analysis, mathematical and computational models provide superb frameworks to test scientific hypothesis. From this perspective, I stick to *Computational Neuroscience* in this work.

On the other hand, engineering the style of information processing in neural systems and developing new computing technologies have been growing recently. *Neuromorphic Engineering* is a new branch that undertakes this mission. In chapter 3 of this thesis I have introduced a new vision sensor technology which imitates the information processing of human retia. I have proposed a novel Neuromorphic algorithm to solve the problem of stereoscopic fusion in these sensors.

In Section 1.1.1 and 1.1.2, I will give an overview of a modern theory in system neuroscience that categorizes the elemental computational units that the nervous system

constantly employs to guide the behavior. Throughout this thesis I adhere to this theory of Brain Computing. In Section 1.2, three main problems of Sensor Fusion are described. And finally, the main contribution of this thesis and the structure of the thesis are elaborated in Section 1.3.

1.1.1 Action-Perception Cycle

All theories that scientists developed during past centuries, generation by generation, began from a very fundamental question: why do we need brain? Within past thousand years, it is argued that this complex organ is encephalized to accommodate the sensation, intelligence, and perhaps the physical basis of intellect (*Al-Farabi* (*872-950*), *René Descartes* (*1596-1650*), and *Baruch Spinoza* (*1632-1677*) supported this idea [Clarke & O'Malley 1996]). As of 18th century, we have realized that this organ functions like a Machine to generate our actions. Which basic functions this machine computes? And how does it compute? In this section I will discuss about the principle functions that brain performs to facilitate the interaction with environment. The second question is addressed in Section 1.1.2.

Survival is the most important goal for any living organism. But, do all animals need a brain to live? Plants can survive without even a single nerve, even though they show a set of very slow reflective behaviors in response to physical stimulations (*e.g.,* light, gravity and temperature). They do not need to move in search of food or a mate. There is also a sea creature, called *sea squirt*[1], which is born with a simple nervous system. This creature can swim until reaching down the ocean and when it settles on some rocks, it starts to digest its brain.

Daniel Wolpert believes that "*the animals need brain to move*" [Wolpert & Ghahramani 2000]. When the sea squirt needs no *movement*, so it does not need a brain. Therefore, it starts to use its brain as a nutritious meal to survive longer. More complex animals naturally demand more complex functions in their lives. Movement is a key ability that enables animals to explore their environment in search of a safe shelter or food to mate or to escape from a predator. All of these actions are associated with necessary goals for survival. So, a comprehensive answer to the question of: why we do need brain?

"*The brain generates a set of goal-directed actions, necessary to maximize our probability of survival* [Trappenberg 2000]".

To maximize the probability of survival, the animal should interact with the environment constantly and through a set of functions (FIGURE 1-3). *Sensation* is the first

[1] Sea squirt is an invertebrate marine animal with potato-shaped body that has some primitive vertebrate features. It is found in all seas, from the intertidal zone to the greatest depths. They commonly reside on pier pilings, ships' hulls, rocks, large seashells, and the backs of large crabs. Some species live individually; others live in groups or colonies.

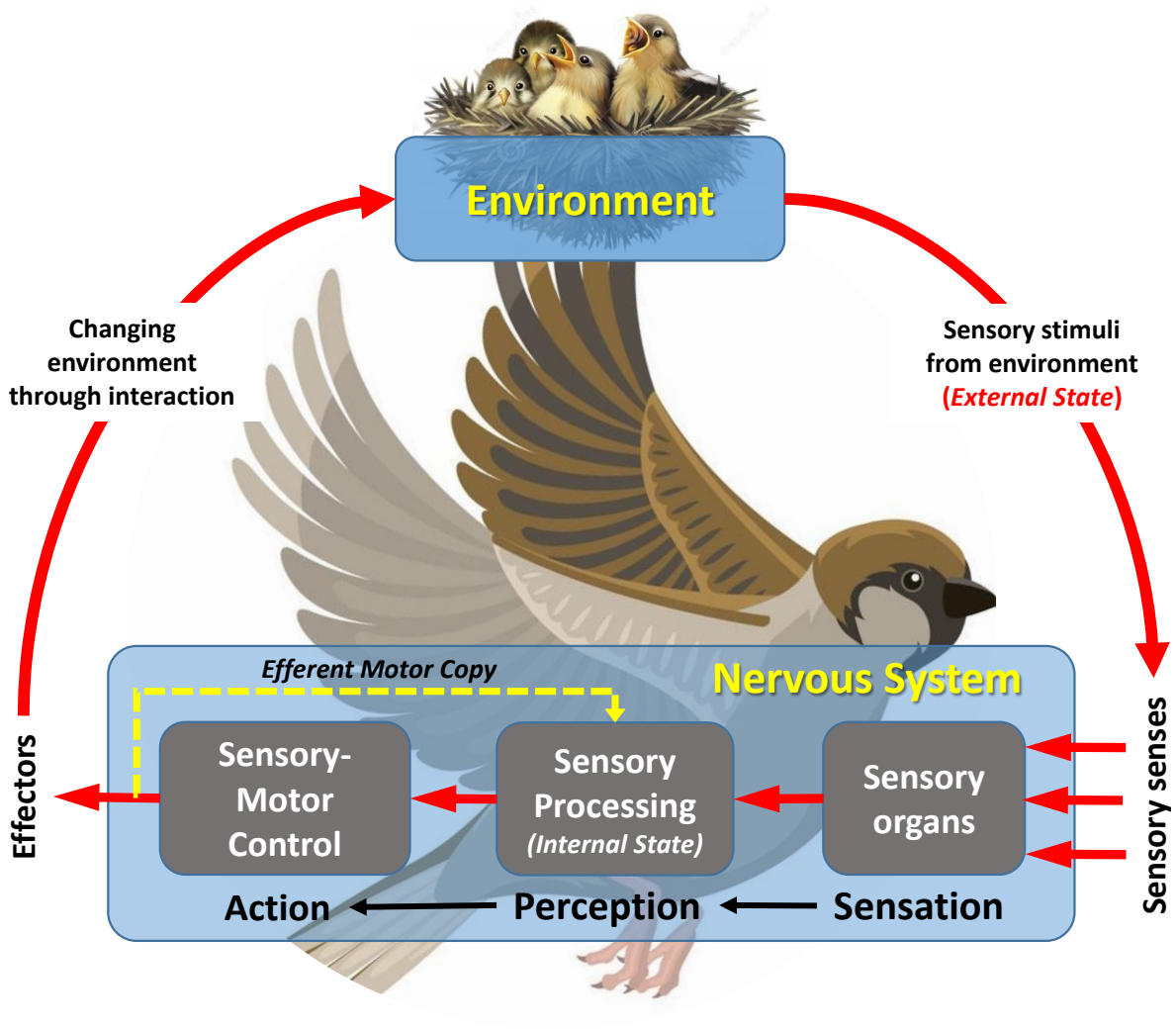


FIGURE 1-3

A demonstration of the animal-environment interaction which is accommodated by Action-Perception cycle. This cycle includes three main elements, Sensation, Perception, and Action-Generation. The goal of the Nervous System is to guide the animal within a safe and optimal trajectory towards her nest. The sensory stimuli of the external world are picked up by sensory organs, transformed into neural activities and delivered into perceptual system, where an internal representation of the sensory world is created. Given an internal percept of the world, the motor system is triggered to generate a sequence of actions and thereby to activate effectors. This will change the state of the animal in the environment (e.g. changing in position) that should be considered by Perceptual system for next cycle.

function by which the physical attributes of the external world are transduced into neural activities – that is often preceded by a transformation of the physical signal in accessory elements of a sensory organ. Then, signals climb up to the thalamus, and thereafter sensory Cortices, where neurons code for an internal map of the physical world (*Sensory*

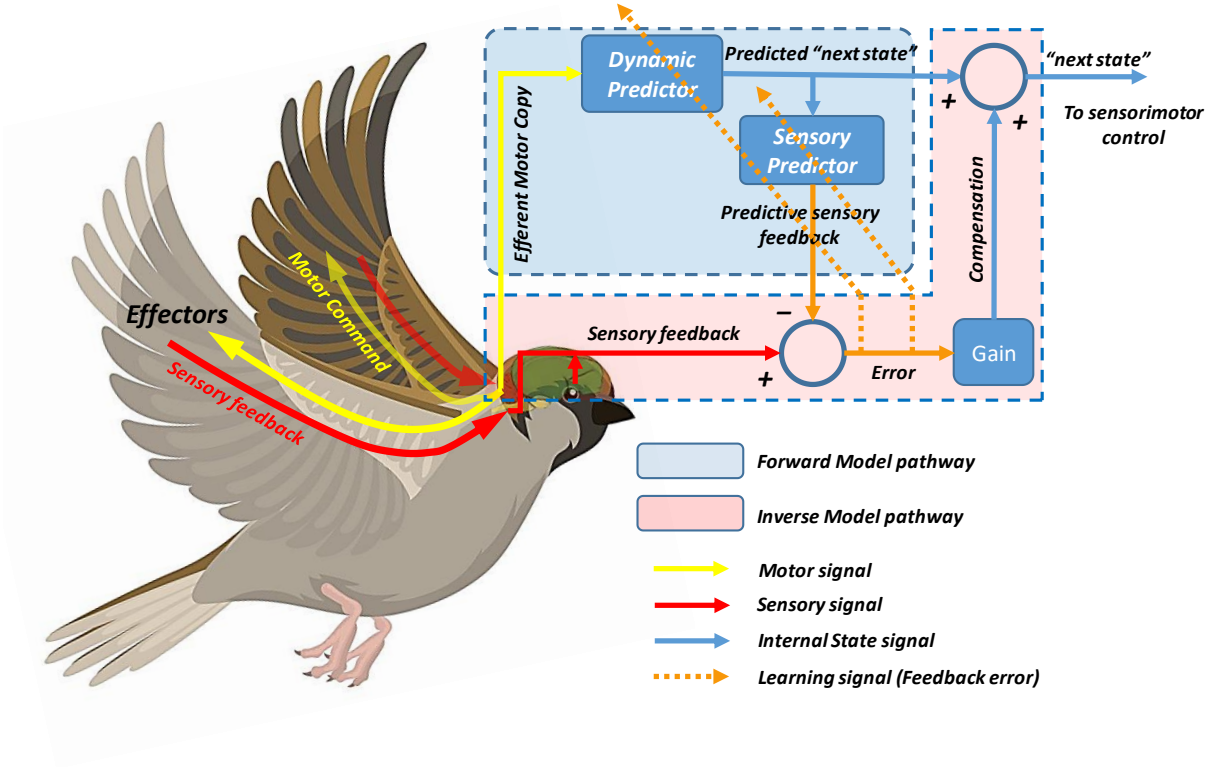


FIGURE 1-4

The representation of Inverse Model and Forward Model in Action-Perception loop. Inverse model pathway transforms the direct sensory feedbacks to a sequence of actions (red pathway). A forward model mediates and compensates this process by a predictive model that predicts the sensory consequence of actions (blue pathway).

Perception) and the body. In accordance with task objective (a higher cognitive concept), which is determined and dictated by higher cortical areas (for example in FIGURE 1-3, the goal is to fly within a safe trajectory toward the nest), and for the given internal state (perceived sensory feedback from the environment including proprioceptive signals), motor system should program and insert a suitable Action. Ultimately, the action will be applied to effectors through cerebellum and spinal interface. The control process of a Motor Action, given desired state, is referred as *Inverse model* (the red pathway is FIGURE 1-4) [Wolpert & Ghahramani 2000]. To successfully guide the animal to reach the goal, it is required to program and initiate a sequence of motor commands in time. These commands will change momentarily the environment. Thus, the internal state which is created by the perceptual system (see FIGURE 1-3), should be quickly updated accordingly, otherwise the animal will be mislocalized and lost. So, the nervous system should always perform the *Action-Perception* process within an Active Cycle. Sometimes this Action-Perception Cycle is referred as *sensorimotor loop* in the literature [Wolpert & Ghahramani 2000].

Feedforward process of Inverse model is not reliable enough to generate the sequence of motor actions, because of two reasons; First, in addition to the sensory transduction in sensory organs, cortical and subcortical sensory processing causes a considerable amount of delay, e.g. about 100ms delay through human visual cortex. Secondly, sensory signals are either infected by noise or are partially observable. Therefore, in practice the motor system cannot only rely on the inverse model. The physiological evidences show that the motor system utilizes a *Predictive Perception,* by which the sensory consequences of the motor action can be internally estimated and used within the control loop. This is another form of *Perception* which is referred as *forward model* (blue pathway in FIGURE 1-4) [Wolpert & Ghahramani 2000].

In other words, forward model uses the efferent copy[2] of the current motor action to predict the internal state at the next time, from which the sensory consequence of the actions can be estimated before any sensory feedback (see FIGURE 1-4). This mechanism enables the nervous system to have a rough estimate of the next state at hand. It is worth to mention that the error signal between the estimated state variables and the original sensory feedback plays an important role in *perceptual learning* (FIGURE 1-4). This signal modifies the animal's belief in the quality of the current action with respect to the goal. A strong coupling between inverse and forward model is actually generating a fast and reliable goal-directed behavior. This form of predictive perception is not exclusively present in the cortical level. It is evident that it also exists at Peripheral Nervous System[3] (PNS) in rabbit [Hosoya et.al 2005], as well as subcortical areas in cat (Lateral Geniculate Nucleus) [Grieve and Sillito 1995]. Since some of the sub-cortical and peripheral neurons are selective for low-level features, early predictive models likely are to directly activate a group of actions that demand a fast response, i.e. visual tracking, saccadic motion [Hogendoorn and Burkitt 2018].

1.1.2 Helmholtzian Brain Computing

In the previous section, I show a block diagram of the computational process that enables the nervous system to interact with the environment. As it is shown, the coupling between sensation, perception, and motor control is essential to rapidly and reliability interact with the real world. Some elemental forms of perception that help the motor-system to generate a goal-directed behavior are introduced. Perception accommodates action-generation, and at the same time the generated action also modifies the perceptual

[2] An efference or efferent signal is a copy of an out flowing movement-producing signal that is generated by the motor system. This copy can be used by perceptual system to predict the consequence of motor actions before it is applied.

[3] The nervous system outside of brain and spinal cord. The main function of the PNS is to connect the CNS to the limbs and organs, essentially serving as a relay between the brain and spinal cord and the rest of the body.

understanding of our world. To summarize the interplay of action-perception, human brain (in general mammals) can be interpreted as a modeler-controller machine.

We control our sensory world while we interact with the physical entities it contains. This process requires an internal model of the sensory events, and a predictive model of the sensory consequences of the actions. To train and to create these explanatory models, humans must explore through the environment to experience the sensory events, and to control them by manipulation. The teaching signal is usually the quality of the actions that the controller applies to reach a goal [Körding & Wolpert 2006]. A good action - in the sense of reaching a goal - will be rewarded and a bad action should be penalized. In machine learning, this learning scheme is called self-supervised learning [Trappenberg 2000] or active learning [Firouzi et.al 2014c] [Firouzi et.al 2014d] [Sagha et.al 2011]. By generalizing the Action-Perception framework, I will take one step forward into a more detailed level of analysis of action-perception cycle and in general brain computation. First, I will give a prevalent definition of perception. And then, in Section 1.1.2.1 it will be theoretically demonstrated how possibly the information is propagated and preserved in the nervous system within a distributed hierarchy. These notions help to understand how the brain computes.

Helmholtz (*1821-1894*) proposed a theory so influential in modern cognitive science that is ruling many developed machine learning algorithms today. His well-known notion on perception is paraphrased as follows [Von Helmholtz 1962]:

"*What we perceive in our sensory world is the conclusion of unconscious inductive*[4] *inference from sensory stimulation, given sensory representation and background knowledge*"

Despite the fact that *conscious awareness* is disregarded in this statement, which gives it a delicate pause, there is no persuasive reason to deny Helmholtzian view to the brain computing [Trappenberg 2000] [Friston 2005] [Boghossian 2014] [Kiefer 2017]. According to the Helmholtz thesis, perception is a subjective inference process in which the current sensory observation is taken as a premise to draw a subjective probability of the potential causes. Our initial belief depends on the previous sensory stimulations that we have experienced through previous action-perception cycles (*background knowledge*). In abstract, perception is a belief modification process and hypothesis testing. When the observer has no idea about a new sensory event, it will draw a rough initial belief in a possible cause associated with previous experiences. The belief might be totally wrong or partially true. Then, by testing the hypothesis about the cause of the current sensory evidence, for instance by taking an action, manipulating an object, or gathering more information about the event, the observer will modify the pervious belief. Belief

[4] The reasoning in which we cannot surely claim the trueness of an argument, we come up with a subjective probability about degree of acceptance of that.

modification can be done either by adding a new belief or erasing the old beliefs [Boghossian 2014]. Since the complexity of the sensory signals varies across different senses, the internal representation of the belief values should be organized within a hierarchy. This functional hierarchy is luckily well-accommodated by *hierarchical structure* of the sensory cortex. More complex features are preserved and represented by higher order regions while low level features are described by early sensory areas. For example, neurons in early visual cortex are sensitive to the angle of alignment, while MT neurons are sensitive to direction of motion. High-order sensory features are sometimes referred as concepts [Trappenberg 2000]. Concepts are in fact concrete elements in the external world, from geometrical shapes and colors, to specific categories of objects, sounds, flavors and qualities. A set of low-level visual features e.g. lines and color, can activate the beliefs in existence of a high-level concept e.g. my show box, see FIGURE 1-5 [Friston 2005]. So, the hierarchical processing along cortical hierarchy is seemingly a key element to create a perceptual belief in sensory world.

When it comes to the belief representation and reasoning, *Probability Theory* provides a rich mathematical framework for formalization [Ernst & Banks 2002] [Shams et.al 2005] [Körding & Wolpert 2006] [Jazayeri & Movshon 2006] [Yang and Shadlen 2007] [Ursino et.al 2011] [Petzschner & Glasauer 2011] [Shams 2012] [Pouget et.al 2013]. On the other hand, it is evident today that human behavior is stochastic. For instance, in a task that one choice of action is rewarded 80% of the time and another 20%, a deterministic system at this scenario always will pick up the first choice. So, there will be no chance for the less probable action to be chosen. Whereas, in real behavior this is not the case. Repeating one identical task, for example reaching task, has always different consequences i.e. arm and hand configurations. This is due to the stochasticity of the external world, neuronal activities, and actuators.

But, are there any systematic advantages for such a stochasticity in brain? To explain the role of noise that can be both destructive and advantages, let us assume a donkey with a deterministic brain[5] that is exposed to two equidistant and identical food rewards. Deterministic world might make the animal suffer the consequences of the indecision and die starving. In fact, earning the wrong food with less reward is better than no food for survival. Similarly, a system without stochasticity and noise might always get stuck into a deadlocked situation. By adding a small amount of noise, we can simply break the symmetry and thereby chose one of the choices, even though it might lead to an unfavorable choice, imprecision or inaccuracy. The main advantage of the neural stochasticity is deploying a computational framework to account for the uncertainty of the sensory signals and accommodating the belief in circuit.

[5] Duns Scotus paradox problem.

To summarize, the following facts shape the governing principles and properties in brain computing:

- Brain is a modeler-controller machine (*Perception-Action*), in which a model of the world needs to be stored and instantiated constantly to activate the motor actions (*Perceptual Inference*).
- The model of the sensory world is represented through a hierarchical distributed architecture in the cortex (*Hierarchical processing*)
- There must a mechanism to consolidate the distributed and hierarchical representation of the world into a unified and coherent from (*Emergent Perception*).
- The beliefs in possible causes of the sensory events are updated by interacting with the environment (*Perceptual Learning*).
- Intrinsic uncertainty in sensory data and motor commands can be internalized by the intrinsic stochasticity of neural activity.

1.1.2.1. Theory of Hierarchical Cortical Responses

In FIGURE 1-5, a formalized framework of the Helmholtzian brain computing is demonstrated where the belief is hierarchically represented. For simplicity, a single modality scenario, i.e. vision, is illustrated. When a physical stimulus that causes a sensory event, C evokes the sensory organ (retina), retina delivers the first form of belief in state variable S^p. The quantity of this variable can be the activity of the neural ensembles in retina. Since the observer is manipulating the environment by taking actions, naturally the belief in sensory state is conditionally related to the previous actions, A^p. So, the uncertainty of the state variable can be formulated according to the following conditional probability function:

$$P(C \mid A^P) \tag{1-1}$$

Given S^p, primary sensory cortex creates the first cortical state variable S^C. While the neural activity of the primary cortical regions depends on the sensory inputs S^p (bottom-up processing), high-level concepts indicated by C', C'' in FIGURE 1-5 can also highlight the relevant information in the primary cortex within a top-down process [Miller 2016]. Considering these two factors, probability distribution of S^C is:

$$P(S^C \mid C', S^P) \tag{1-2}$$

Concepts, which are in fact high-level explanatory variables, are represented hierarchically. So, higher-order concepts that are either evoked by sensory inputs or a higher cortical level can also change the expectation of the concepts in the lower cortical areas, C′ is a low-level concept that can be described by as follows:

$$P(C' \mid S^C, C'') \tag{1-3}$$

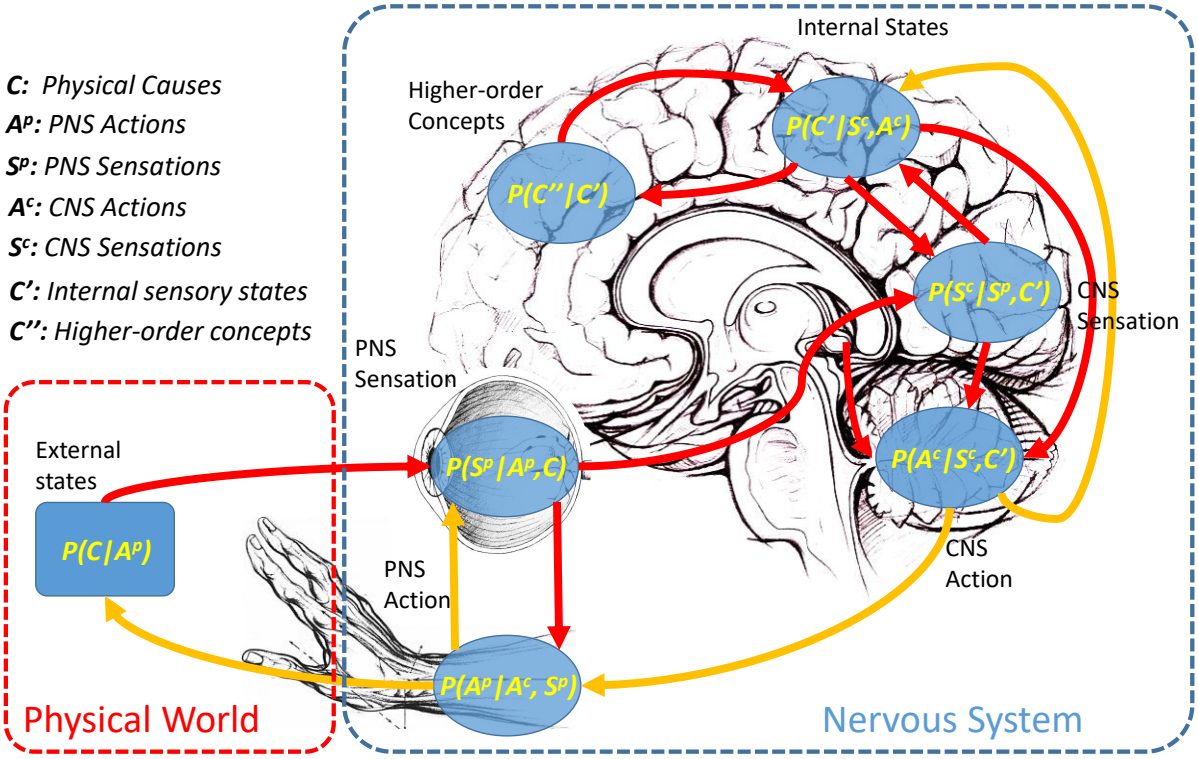


FIGURE 1-5

A schematic formalization of Helmholtzian brain. A distributed network of probabilistic nodes preserves the belief in sensory state variables at certain levels of complexity. High level features (or state variables) are called Concepts. A single physical stimulus C (or a cause) evokes the sensory organs (peripheral sensory system), and thereafter activates the sensory cortex (S^P and S^C vectors respectively). A bottom-up stream of information creates a hierarchical internal map of the world, while top-down information will mediate the perception and thus action. The high-level concepts (C'' and C') are activated by low-level concepts (C') and sensory state (S^C), while sensory states are also effected by high-level concepts. This form of information exchange is synonymous to the reciprocal connectivity of cortical areas. The main idea of the picture is taken from *Friston's* theory of cortical responses [Trappenberg 2000] [Friston 2005]. This network resembles the Belief-Propagation-Network in Machine Learning.

Ultimately, given the internal cortical state variables S^C and the activated concepts C', CNS should map the perceived state of the world into a set of appropriate actions A^C. Motor signals programmed and generated in CNS need to pass through spinal interface to reach effectors located in PNS. In this pathway the signals are infected by an intrinsic noise. To include motor noise, PNS action A^P is defined within a distinct distribution function:

$$P(A^P \mid A^C, S^P) \tag{1-4}$$

The schematic representation of FIGURE 1-5 is sometimes called *Deep Belief Network*. The term "deep" constitutes the hierarchical architecture of the network, and "belief network" represents momentary expectations of state variables at certain level. Some of the nodes are connected reciprocally showing the conditional connectivity of variables. Synonymously, the sensory cortex is organized hierarchically, and cortical connectivity is reciprocal. Forward and backward connections show a functional asymmetry so that forward connections carry driving signals, and backward connections are both driving and modulatory [Miller 2016].

As I noted, the neural activity in the nervous system is governed by stochasticity. For instance, in a wide range of cortical neurons, the response of the neuron for an identical stimulus fluctuates according to a Poisson-like distribution [Ursino et.al 2011]. But, the fundamental question is how probabilistic Poisson neurons can mechanistically compute probability distribution functions? On the other hand, as you can see in FIGURE 1-5, the nervous system requires to perform marginalization over specific set of variables in the information pathway. In Chapter 5, I will show how a linear combination of *Poisson neurons* can encode a random variable with a Gaussian-like probability distribution function. This neural coding scheme is called *Probabilistic Population Code* that is used in this thesis [Jazayeri & Movshon 2006] [Pouget et.al 2013].

1.2 Multisensory and Sensorimotor Integration

1.2.1 Multisensory Integration in perceptual inference

In FIGURE 1-5, a general scheme of the Helmholtzian brain computing is demonstrated that can produce the sensation-perception-action cycle in a single modality i.e. vision. However, the world is composed of different attributes that should be captured by the perceptual system and combined into a coherent representation (*emergent property*). For example, in FIGURE 1-3, the bird receives acoustic, visual, and geographical signals[6] from the environment, in addition to the proprioceptive cues from her body to create a spatial map and to generate a sequence of actions accordingly. Moreover, the bottom-up stream of information (see FIGURE 1-5) will activate the higher order concepts that are mostly composed of multiple attributes across different modalities. For instance, a picture of a dog that activates some regions of visual pathway is associated with the sound of barking that activates some areas of the auditory pathway.

[6] The geomagnetic field can provide animals with two kinds of information: The magnetic vector provides directional information and can be used as an internal compass. While the total intensity and/or inclination provide information on the position used for navigational processes or acting as triggers. There are several beautiful experiments that have studied birds (European Robins, Chicken) and Turtles, and have shown that the nervous system of these animals are able to pick up these information for navigation. In birds, the magneto-receptors are located in their right eyes (direction preferred like compass) and the intensity sensitive receptors are located in their upper beak.

As is discussed already, the sensory and motor signals are infected by uncertainty, either by external or internal noise. For instance, imagine a bird that must navigate through a foggy field. In this case, the animal cannot rely on the sense of vision, and the perceptual system should reduce the contribution of visual information compared with other senses (see FIGURE 1-3). Therefore, the perceptual system should use a mechanism to deal with the varying quality of signals and to perform a flexible form of sensory combination in a cluttered environment. Moreover, having multiple sources of information at hand enables the perceptual system to reduce the pitfall of the twisted sensory nodes[7] and to identify the possible defects. In other words, if a single sensory organ suffers from deprivation or deficits for any reasons, there are alternative sources available to compensate and to calibrate the faulty node. This is another advantage of the multisensory perception.

The process of combining different form of attributes and physical descriptors of an environmental event (see FIGURE 1-5), which is meant to be perceived as reliable and accurate as possible, is called *Sensor Fusion* or *Multisensory Integration* (sometimes referred as *Cue Integration*[8] in literature). The process of perception is highly multimodal, because the world is intrinsically multimodal and carrying multiple forms of information. However, information integration can also take place in a single modality. For instance, to form a consistent percept of visual depth, the visual system combines retinal disparity[9] with geometrical information and statistical characteristics of the visual scene [Banks et.al 2011]. Or at early visual system, the action potentials of retinal ganglion cells are combined in striate cortex, so that the single neurons are spatially registered to encode a specific angle of orientation.

Another example is emotion recognition, by which the emotional state of a speaker can be recognized by combining several auditory features within a hierarchical processing [Sezgin et.al 2012]. This process of combining sensory information within single modality is referred as *Unimodal Sensor Fusion*. Most of the computational principles, either in functional or neural level, that govern the process of Unimodal Sensor Fusion, are basically similar to those that shape Multimodal Integration. On the functional level both cross-modal and within-modal integration can be modeled by a single formalism. A particular successful and powerful framework is Bayesian Integration [Ernst & Di Luca 2010] [Körding & Wolpert 2006] [Ursino et.al 2011] [Banks et.al 2011] [Landy et.al 2011] [Ernst and Bülthoff 2004] [Alias & Burr 2003] [Kersten et.al 2004] [Bisley 2011] [Yang &

[7] As we follow the theory of Helmholtzian brain computing, we model each sensory signal as a node described by a conditional probability. Each node can generate a new belief value and propagate it within the network (see FIGURE 1-5)

[8] Sensor Fusion is mostly used in Engineering, and Multisensory Cue Integration is often used in biology and psychology.

[9] In Chapter 3 we will describe the concept of retinal disparity as an important visual cue for depth perception.

Shadlen 2007] [Fetsch et.al 2011].In Section 1.2.2 and 1.2.3 and 1.2.4 I will briefly describe three problems in *Sensor Fusion* that should be solved by the perceptual system.

1.2.2 Reliability, Optimality, and Accuracy

1.2.2.1. Reliability and optimal estimation:

The underlying mechanisms of Multisensory Integration in the brain is context-dependent [Boyle et.al 2017]. In the context of action-perception, the main functional role is to minimize the negative consequences of noise (or equivalently maximizing the reliability) and to cancel out the systematic sensorimotor inaccuracy. Reliability reflects the quality of signal or the amount of information that sensory node carries about physical state. The best way to quantify the reliability is to measure the fluctuation of the sensory node (or the frequency of observer response) in response to an identical physical stimulus. The real physical value of the stimulus is not directly available and needs to be estimated from current noisy observation. For instance, in FIGURE 1-6 visual and acoustic responses given an identical stimulus fluctuate around single values (Maximum Likelihood value) $\hat{S}_V$ and $\hat{S}_A$ respectively. In other words, given the real sensory stimulus S, the probability distribution functions represent how likely the current sensory observation can be generated (the likelihood probability of the current sensory observation). This fluctuation can be best reflected by the variance of the likelihood functions $L_i(S)$, and that are reversely related to the reliability of the sensory signals. So, generally we define the reliability as the inverse of variance. In FIGURE 1-6, σ_A is twice greater than σ_V, or equivalently $\hat{S}_A$ fluctuates less than $\hat{S}_V$ around their mean values ($\hat{S}_i = max\{P(S_i|\mathrm{S})\}$; in normal distribution, it is equal to mean value). That shows a higher reliability for visual estimate compared with acoustic estimate[10]. Now the question is *how to combine sensory observations to optimally minimize the fluctuation of cross-modal estimate*(σ_{AV})*?* Naturally, the transformation of unimodal signals to a single multimodal estimate (assuming that sensory nodes are representing a single physical value S) must give a higher weight to more reliable signal. Let us assume we employ a simple linear combination strategy, where the integrated multisensory estimate $\hat{S}_{AV}$, is a weighted average of individual sensory estimates:

$$\hat{S}_{AV}^{opt} = w_A\hat{S}_A + w_V\hat{S}_V; \quad w_A + w_V = 1 \tag{1-5}$$

Intuitively, the best candidate for w_A and w_A is the relative reliability of two signals which are reversely proportional to their corresponding variance values:

[10] The current estimate ($\hat{S}$) is equal to the current sensory observation (or signal). Given an identical sensory stimulus (S), if we repeat that sensory stimulus and collect subject responses over time, the likelihood function will be emerged. Because the response is fluctuating as a result of sensory noise (see red and green curves in **FIGURE 1-6**).

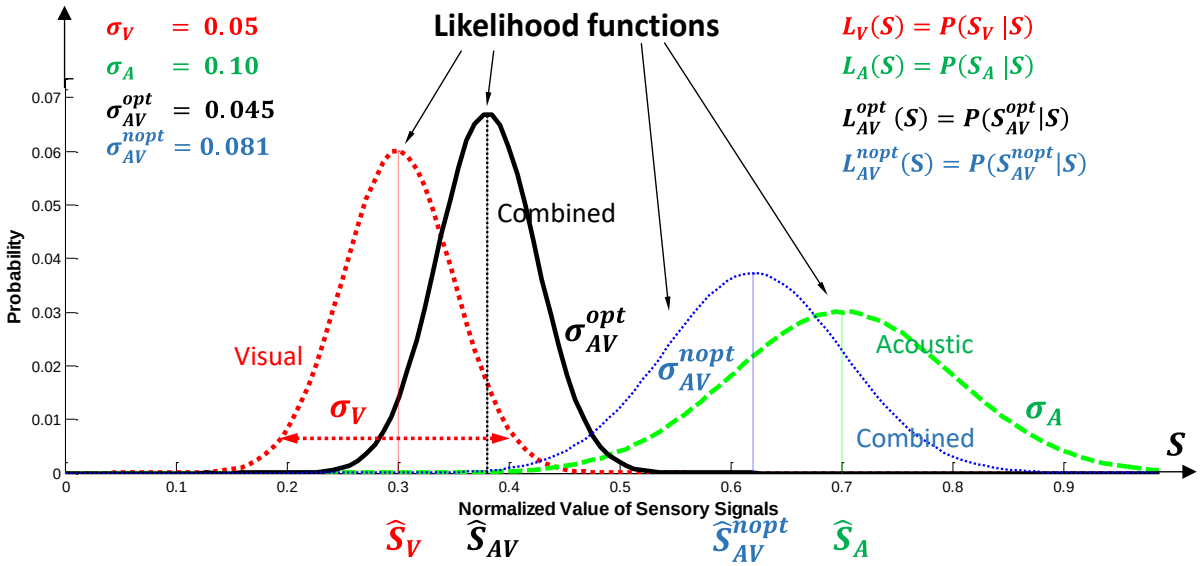


FIGURE 1-6

A Schematic representation of sensory likelihood functions (Visual and Acoustic in red and green respectively), and likelihood of combined estimates, $\hat{S}_{AV}^{nopt}$, $\hat{S}_{AV}$ (Black and dashed-line Blue). The black-colored Gaussian curve shows the outcome of an "*optimal*" estimate, and the dashed-blue curve represents a "*non-optimal*" combination. The quality of each signal is reversely proportional to their respective variance (the spread of Gaussian functions).

$$w_i = \frac{r_i}{\sum_i r_i};\ r_i = \frac{1}{\sigma_i^2} \tag{1-6}$$

The linear combination of two Gaussian random variables is another random variable with a Gaussian distribution. In FIGURE 1-6, the black Gaussian distribution represents the probability distribution function of the linear combination of two single-modality estimates (Equation 1-5 and 1-6). As is depicted in FIGURE 1-6, the variance of the multisensory estimate σ_{AV}^2 is reduced compared with the variance of each single-modal estimates (σ_A^2, σ_V^2). This demonstrates the benefit of multisensory combination in reducing uncertainty. Now if we flip the assigned weights in equation 1-5, in such a way that the auditory estimate holds higher contribution, the likelihood distribution of multisensory combination (blue curve in FIGURE 1-6) will be shifted toward acoustic distribution (green curve in FIGURE 1-6). In this case, the variance of the combined estimate is also closer to the acoustic likelihood and increased drastically compared with the previous scenario in which w_i is proportional to the respective sensory reliability. Similarly, if we increase the weight of visual estimate (a value greater than w_v in equation 1-6), the likelihood curve of the combined signal will be shifted toward visual likelihood function, and thereby, the σ_{AV} will rise slightly compared with σ_{AV}^{opt}. Thus, the *optimal strategy* to combine the sensory signals is to follow equation 1-5 and 1-6 and to weight signals according to the relative variances. Along with the problem of *Optimality,*

assigning a suitable weight to the sensory signal is known as *validity problem* in sensor fusion. In **Chapter 2,** we will prove why MLE is an optimal computational strategy under certain circumstances.

Typical models of multisensory integration assume a normally distributed independent source of noise within single modalities. This assumption in general is likely to be true as the governing neural processing for each modality is independent [Landy et.al 2011] [Ernst & Bülthoff 2004]. In FIGURE 1-6, it is assumed that sensory observations are infected by an independent source of Gaussian noise. Equation 1-5 and 1-6 are referred as Maximum Likelihood Estimates as the best estimate - either within-modal or cross-modal - is the one that maximizes the corresponding likelihood function. MLE is considered as the standard model of sensor fusion. There is a large body of psychophysical and neurophysiological studies that corroborates the fact that the nervous system employs MLE in a wide range of multimodal perceptual tasks (e.g. visual-haptic size discrimination task [Ernst & Bülthoff 2004], audio-visual localization [Alias & Burr 2003], and object recognition [Kersten et.al 2004]).

1.2.2.2. Accuracy and Systematic Bias:

MLE is an optimal strategy for sensor fusion only under certain constraints:

1. The sources of sensory noise must be statistically independent and uncorrelated.
2. noise is normally distributed.
3. The single-modal sensory experiences are uniformly distributed[11].
4. The sensory estimates must be unbiased and accurate.

There are situations that might not hold at least one of these constraints, and thereby, the standard form of sensor fusion (reliability-based weighted averaging) is not an optimal combination strategy. Sometimes, it is possible that one of the sensory inputs provides highly reliable information, but its mean value deviates from the real physical value. In this case, the combination will be error prone, because we give a higher credit to the most reliable but biased signal. As a result of that, the multisensory estimate will be drastically biased from physical value, and thereby suboptimal. For instance, in FIGURE 1-6, the likelihood function of the multisensory fused signal (black curve) is biased toward a visual estimate which is the more reliable signal (red curve), whereas its noise content is reduced compared with both single-modal estimates (the spread of black curve is reduced compared with red and green curves).

[11] Prior probability indicates the frequency and probability of a sensory stimulus within the sensory space, while likelihood is govern by intrinsic noise and indicates the likelihood of current sensory observation given an individual sensory stimulus.

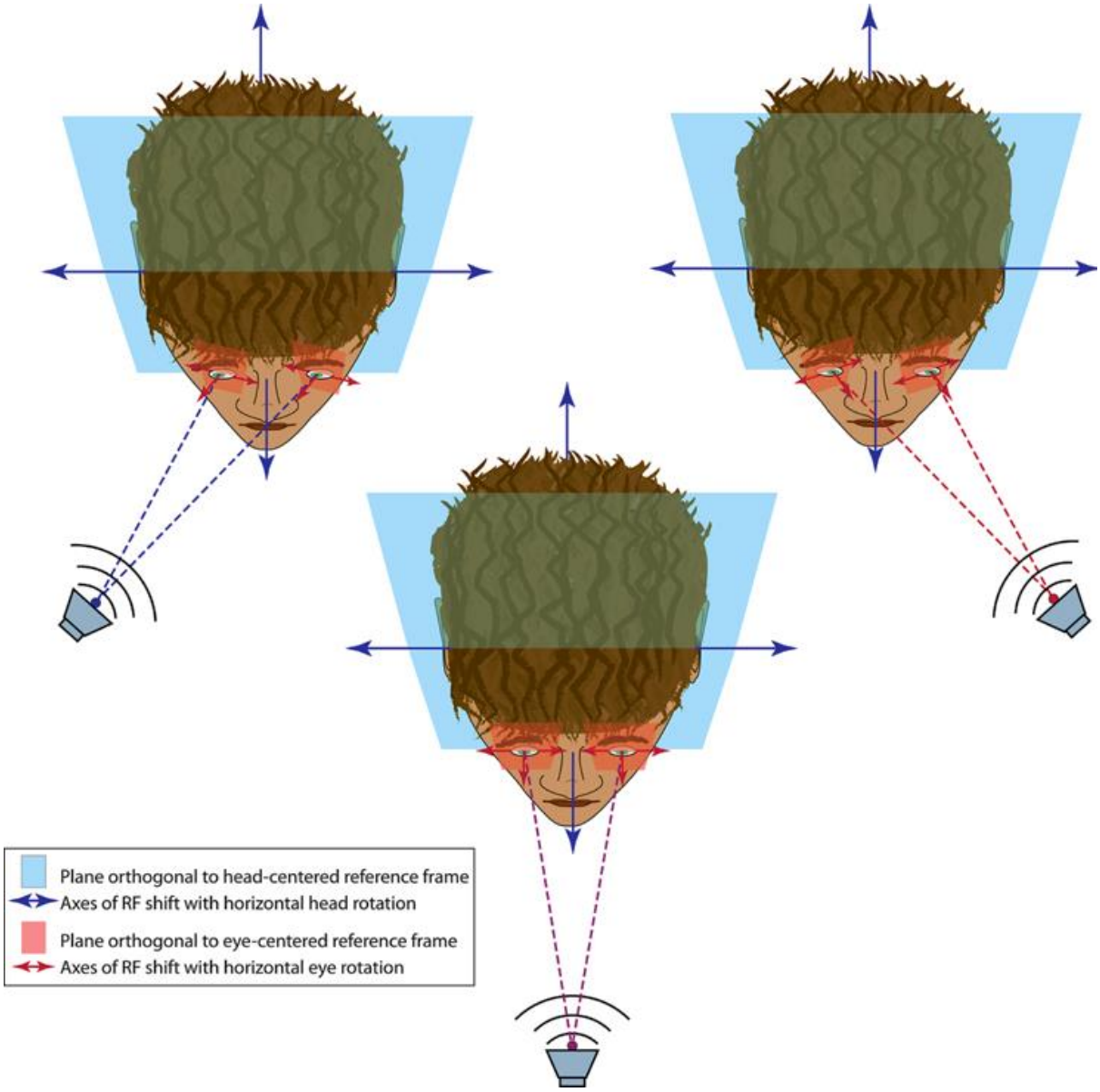
To benefit from the advantage of fusion (noise reduction) and to avoid its cost of introducing bias to the final estimate, the nervous system should infer the accuracy of the signal sources and re-calibrate or disconnect the sensory defects. This task is very difficult for the nervous system since because unlike the reliability of the signal, bias cannot be directly determined from the sensory estimates. Nevertheless, the *problem of validity* must also take into account the persistently biased sensory node. In next chapter, I will analytically explain how possibly the nervous system estimates this information and how possibly conducts re-calibration to avoid both noise-driven and bias-driven inaccuracy (Section 2.3.3).

In addition to assigning suitable weights to the sensory signals with respect to signal variation and bias, an optimal strategy must use all pieces of available information to minimize *Mean Square Error*. The statistical structure of the sensory world is another piece of information that the nervous system incorporates in a multisensory combination, either cross-modal or within-modal integration. For example, in natural visual stimuli, the frequency of observing vertical or horizontal edges are higher than intermediate angles [Lee and Yuille 2007] [Girshick et.al 2011].

Therefore, the expectation of visual perceptual system is internally biased toward vertically or horizontally aligned features. This statistical property is called *sensory prior* which is best formalized by a probability distribution function called *prior distribution*. This probability function also indicates the joint probability of multimodal sensory events, and thereby, models multisensory correlation. To include the prior in multisensory fusion and to generalize an optimal form of combination, Bayesian decision theory provides a unique mathematical framework. Within this framework, the probability of the sensory stimulus, given current noisy sensory observations (posterior probability), is proportional to prior probability times likelihood probability. This scheme of integration is in fact anchored in Helmholtzian theory (see Section 1.1.2), where the perceptual system is defined as an inference machine to combine sensory representation (*Sensory Likelihood*) and background knowledge (*Sensory Prior*). In Chapter 2, we will see a comprehensive mathematical description of Bayesian Integration method.

1.2.3 The Problem of Reference Alignment

The sensory signals from different modalities encode physical features with different dynamic range and properties. To combine the acoustic location of a singing bird with its visual location, the signals must be aligned into a common-frame of reference. On the other hand, the relationship between acoustic and visual locations can be possibly changed by eye or head motion. Human ears are a pair of head-mounted organs whose coordinate frame of reference is head-centered. Whereas, human eyes moves in vertical and horizontal directions and thereby the retinal coordinate can be shifted by ocular motion. Another example is the angular position of the joint-ankles in free 3D space body



By courtesy, taken from [Gruters et.al 2012].

FIGURE 1-7

A demonstration of reference-frame-alignment problem in Multisensory Integration. The blue frame represents head-centered coordinate within which auditory space is mapped (azimuth and elevation angels). The red frame represents eye-centered visual frame of reference. At central panel, when an audiovisual target is straight ahead, the auditory and visual reference frames coincide and are perfectly aligned; and both are perpendicular to the audiovisual information source (speaker). When the target is moved eccentrically (upper panels), the head-centered and eye-centered reference frames are no longer parallel, with the eye-centered reference frame having rotated around a vertical axis (for horizontal eye movements). In this example, the eye-centered reference frame is still perpendicular to the audiovisual target, but the head-centered reference frame is not.

define a non-linear coordinate system that can be mapped into visual depth which is defined in body-centered coordinate. So, the nervous system should give the perceptual system a mechanism that accounts for aligning the frame of references within different modality coordinates. In FIGURE 1-7 a single example of the alignment problem in audio-visual perception is illustrated, where the eye-centered coordinate must be aligned with head-centered coordinate by using ocular proprioceptive cue (position of eye). *Reference alignment* is a key process that exclusively takes place within multimodal domain. Nowadays, there are multitude of neurophysiological studies in human, primates, and cats that have identified the cortical regions - mostly parietal regions i.e. *LIP*[12] and *PPC*[13] that perform reference alignment in audio-visual, vestibulo-ocular, and ocular-visual space [Cohen & Andersen 2002] [Brostek et.al 2015] and subcortical areas IC[14] and SC[15] [Wallace & Stein 1997] [Stanford et.al 2005] [Boyle et.al 2017]. Pouget and colleagues have theoretically investigated how possibly cortical circuits can implement this process by using basis functions and probabilistic population codes. This model can remarkably describe many aspects of reference-alignment in human [Pouget & Sejnowski 1997] [Pouget & Snyder 2000] [Avillac et.al 2005].

1.2.4 The Problem of Credit-Assignment

As is demonstrated in FIGURE 1-6, under specific assumptions the Perception highly benefits from multisensory combination. For instance, a linear combination of two sensory signals is beneficial just in case the weights of linear combination are reversely proportional to the variance of the respective sensory nodes (see equation 1-5 and Section 1.2.2.2). Another constraint that guarantees the optimality of this approach is that the sensory signals originate from a common source. The combination of signals under this assumption is called forced-fusion. But In the real world, we are constantly surrounded by multiple objects and therefore by multiple sources of information. If the signals come from different sources, a reliability-based linear combination is not an optimal strategy anymore and leads to a systematic bias. On the other hand, it is not rational to combine different attributes of two distinct sensory events. Along with the problem of validity (determining the reliability of the sensory nodes), and to generalize the problem of optimality, the perceptual system should take into account all possible scenarios that can happen in a multisensory task. Bayesian Decision Theory has bestowed a revolutionary methodology to model the underlying process of multisensory and sensorimotor combination in a wide range of perceptual decision tasks. In Chapter 2 I will give a comprehensive survey of this approach. Besides forced-fusion, segregating information conveyed by separate sources plays a vital role in multisensory perception. Taking into

[12] Lateral Intraparietal Cortex
[13] Posterior Parietal Cortex
[14] Inferior Colliculus
[15] Superior Colliculus

account all possible existing hypotheses in the environment and the problem of whether to fuse information or not, involves a probabilistic process called *Causal Inference*. This problem is posed in a higher level of cognitive complexity compared with linear fusion. The problem of causality sometimes is referred as *Credit-Assignment* problem in the literature [Berger 2006]. It is not yet fully understood where and through which mechanism Causal Inference is emerged in the sensory cortex. This problem is less investigated as compared with forced-fusion and Bayesian Integration.

1.3 Thesis Structure and Contributions

In Section 1.2 the main three problems of Sensor Fusion are defined that must be considered by the perceptual system; the problem of optimality and validity, reference alignment problem, and the problem of credit-assignment. In Chapter 2, a thorough literature review of the research works and the state-of-the-art models that deal with the problems of sensor fusion in the context of multisensory perception is given. The models that are more relevant to this research work are reviewed. Most of these algorithms are designed to describe a specific property of the multisensory integration in human, under specific assumptions. First the mechanisms of sensor fusion according to the level of data-fusion is categorized. Then, two different approaches are explained and categorized: deterministic and probabilistic models. In deterministic approaches, the correlation between sensory nodes that dictates a relaxation dynamic, or a mutual prediction process, is described using a manifold. In probabilistic approaches - mostly anchored in Bayesian theory - the beliefs in possible hypothesis regarding the sensory events and a prior belief are described and combined by using conditional probability functions. We have introduced the main theoretical works that have investigated the cognitive signature of multisensory integration using model-based approaches.

In Chapter 3, we introduce the problem of stereoscopic fusion as one of most well-known problems in vision. The advantages of the style of information processing in neural systems have attracted engineers to develop more efficient sensors and processors [Soman et.al 2016]. Neuromorphic vision is a growing technology that focuses on engineering the neural functions of human retina. In contrast to conventional vision sensors, human retina encodes visual information using asynchronously generated spikes rather than clocked-frames. This makes the problem of stereoscopic fusion more complicated since the correlation between incoming spikes should be computed in an asynchronous way, or what we call it: in an event-based fashion. Biologically-inspired event-driven silicon retinas, so called dynamic vision sensors (DVS) imitate the functionality of human retina, and thereby allow efficient solutions for various visual perception tasks, e.g. surveillance, tracking, or motion detection. Similar to retinal photoreceptors, any perceived light intensity change in the DVS generates an event at the

corresponding pixel. The DVS thereby emits a stream of spatiotemporal events to encode visually perceived objects that in contrast to conventional frame-based cameras, is largely free of redundant background information [Lichtsteiner et.al 2008]. The DVS offers multiple additional advantages, but requires the development of radically new asynchronous, event-based information processing algorithms. In Chapter 3 I have proposed a novel fully event-based disparity matching algorithm using dynamic cooperative neural network [Mar 2010] [Firouzi and Conradt 2016]. In this network, the interaction between cooperative cells applies cross-disparity uniqueness-constraints and within-disparity continuity-constraints, to asynchronously extract disparity for each new event, without any need of framing individual events. We have investigated the algorithm's performance in several experiments; our results demonstrate smooth disparity maps computed in a purely event-based manner, even in the scenes with a complicated temporally-overlapping stimulus. This work is one of the first successful attempts to solve the problem of stereoscopic fusion in event-based vision sensors.

In Section 4.1, using the theory of Dynamic Neural Field, and by extending the basic model of visual attention proposed by [Rougier 2006], we have proposed a hierarchical recurrent neural model that demonstrates the cognitive advantages of the predictive perception (see FIGURE 1-4). In this network, a rough estimation of the visual motion is computed using a recurrent neural network (motion field), then, this network provides a top-down feedback to early visual areas (focus field). This feedback connection in fact adds an extra evidence regarding the location of the target at the next time step. When the sensory evidence is provided, the overlapping area of the sensory signal and predictive signal will highlight the location of the attended object. This network has theoretically demonstrated how visual motion-cue can possibly be integrated within a visual spatial-map and thereby that guide the behavior (visual overt tracking) in favor of an attentional goal. The performance of this approach is evaluated using artificial data in an extremely noisy situation, in presence of realistic salient distractors, and in a realistic collision scenario. The prominent advantage of cue combination in this network is demonstrated in collision scenarios where the target collides with another salient object. Most of the conventional saliency-based models fail to capture the target in this case [Itti & Koch 2001] [Rougier 2006] [Rougier & Vitay 2006] [Rougier & Vitay 2011]. But, cueing the attended target with a motion signal helps the observer to keep tracking the target even in presence of a salient distractor. The data recorded from a Dynamic Vision Sensor is used to assess the performance of the network. The main functionality of this predictive model is similar to that of described in FIGURE 1-4.

In Section 1.2.2, it is shown how MLE provides a simple solution for the problem of validity and optimality in a Gaussian process. There are a vast body of research that have investigated the behavioral correlates of linear MLE in Audio-Visual localization [Alias and Burr 2004a] [Wallace et.al 2004], Audio-Visual synchrony [Shams et.al 2005], visual-

tactile size discrimination [Ernst & Bülthoff 2004] and vestibulo-Ocular heading estimation [Fetch et.al 2011] tasks. Fetch and colleagues [Fetch et.al 2011] studied monkeys that perform heading-estimation using visual-motion and vestibular signals. Strikingly they found that the monkeys are able to combine the cues according to the varying values of the signal reliability [Fetch et.al 2009]. However, Fetch and colleagues left the question open what is the neural correlate of this computation? In addition to that, it is still unclear how the statistical properties of the sensory modalities, e.g. reliability, prior, and probability distribution are coded in neural circuits. A straightforward approach is that a dedicated area in the sensory cortex or thalamus receives the uni-sensory signals and combines them into a single percept. Most of the neural models of MLE and Bayesian integration follow this approach [Ma et.al 2006] [Alvarado et.al 2007] [Magosso et.al 2008] [Ursino et.al 2011] [Ursino et.al 2014]. However, this feedforward architecture is not the only way that neural circuits can possibly perform integration.

On the other hand, this hypothesis is not in line with recent experimental findings in which the interconnection between multiple poly-sensory regions in sensory cortex facilitates multisensory integration [Chen et.al 2013]. In Section 4.2, using plausible neural principles and attractor dynamics, we have proposed a neural model in which multiple neural ensembles are mutually connected and receive uni-sensory signals through feedforward connections. A modified version of this framework enables a reasonable degree of flexibility to train an arbitrary relation function, and thus is capable to perform relation satisfaction and reference-alignment. Cook and colleagues proposed an unsupervised framework of relation learning between two interacting populations of neurons, which allows the network to learn arbitrary relations between two encoded variables [Cook et.al 2010].

However, a flexible computational framework which could learn relationships between cues rather than using fixed networks is still addressed as a challenge, especially in the presence of higher order modalities [Cook et.al 2010]. In contrast to the common approach of converging zone (explained in this paragraph), there is no single exclusive multisensory area that accommodates the unified percept. Rather, the attractor dynamics between interacting areas preserves the combined signal across multiple pathways. This style of information processing resembles to that of theorized by the theory of cortical responses and explained in FIGURE 1-5. Another issue in multi-sensory integration which is less investigated is how to encode and learn reliability of cues into spatially registered form of neural activities. It is also important to note that the reliability of the sensory signals is not uniformly distributed. For instance, the location of visual stimuli near fovea is more reliable and identifiable than periphery ones. This circuit can perform reliability-based sensor fusion by means of attractor dynamics in which the relative reliability of the sensory cues are encoded within the gain of neural activities.

In the proposed attractor neural network each sensory cue is encoded by a single population of neurons that are laterally interconnected. Each population is also mutually connected to other populations. We have investigated a sensory convergence experiment and it is shown how modulating the neural activity according to the relative reliability of encoded cues can bias the dynamics of the network in favor of more reliable cue. As a result of this modulation, the dynamic of the attractors would be in favor of more reliable cue and the relaxed condition (decoded value of combined sensory estimate) is proportional to an optimal MLE estimator. This model is evaluated in a tri-modal heading estimation experiment using an omnidirectional mobile robot [Firouzi et.al 2014b]. We have compared the outcome of the network with MLE and it is shown that the network can realize a near optimal solution for reliability-based multi-sensory cue integration. We show how Gain Field Modulation (GFM) can modify the dynamical behavior of the network in favor of more reliable cue. Gain modulation is a well-known mechanism that brain uses to highlight information under specific internal or external constraints, e.g., attention-based modulation of striate cortex by higher cortical feedbacks [Bisley 2011]. This mechanism is also observed in monkey and human, in which the varying quality of sensory stimuli modulates the neural activity of the sensory selective neurons in visual cortex [Yang and Shadlen 2007] [Fetsch et.al 2011] [Boyle et.al 2017].

The problem of credit-assignment in multisensory perception is less explored as compared to forced-fusion. This computational process can be described as an inference process, since the perceptual system must compute a belief in the existing hypothesis regarding the cause of the sensory events. The less investigated problem in multisensory research is understanding the underlying neural mechanisms of the multisensory causal perception in cortex. There are very few research works that studied the behavior of human subjects in a multiple hypothesis scenarios [Wallace 2004] [Körding et.al 2007] [Shams 2012] [Rohe & Noppeney 2015]. Some few theoretical works also tried to shed some lights to understand the governing neural principles of Multisensory Causal Integration [Weisswange et.al 2011] [Ma & Rahmati 2013].

Most of these models are either non-plausible, e.g. [Ma and Rahmati 2013], or are incapable of describing the main characteristics of behavioral data [Weisswange et.al 2011]. In Chapter 5, we have reformulated this problem in a way that it can be mapped into a plausible hierarchical neural circuit. A recent fMRI study on human subjects performing audio-visual localization, identified the cortical regions that are involved in Multisensory Causal perception. They show that this process is likely emerged by a hierarchical distributed circuit along parietal and early sensory cortices [Rohe & Noppeney 2015]. The architecture of the proposed circuit resembles the functional hierarchy identified by [Rohe & Noppeney 2015]. The proposed model can successfully reproduce the psychophysical data in audio-visual perceptual decision. When auditory and visual stimuli are largely spatially inconsistent, the fusion pathway is inhibited with

a higher probability, implying the fact that signals are likely caused by separate sources. In Chapter 5 we have described the detail mechanics of this model. The results are also demonstrated in Section 5.4. The results support the notion of de-centralized multisensory integration which is the central hypothesis of this thesis.

Chapter 2

Computational Models of Multisensory Integration, from Perception to Action

"There are no incurable diseases — only the lack of will. There are no worthless herbs — only the lack of knowledge"

— Avicenna (980 – 1037 AD)

2.1 Introduction

Sensor Fusion is not exclusively exposed and studied in Brain Research. Early algorithms in sensor fusion are developed for military applications. This term is primarily used in Computer Science and Information Theory to address those algorithms that combine multiple sources of information to improve the quality of information. Hence many techniques are widely developed in multitude of contexts, including control theory, robotics, signal communication, signal detection and classification, target identification and tracking, image processing and remote sensing, medical imaging, etc. [Hall et al. 2009]. As we briefly discussed in Chapter 1, the main advantage of using multiple sources of information is to reduce the intrinsic uncertainty of the environment, detect sensory defects, and improve accuracy for more system reliability that cannot be achieved by using a single source of information. In general, this procedure is called Multisensory Integration. The second benefit of Sensor Fusion is to extend the range of sensory measurement. Each sensor may measure a specific range of information, so it is necessary to combine different ones to gain a wider range of observation. To be more literal, this type of multisensory integration process is called Multisensory Combination [Hall et al. 2009].

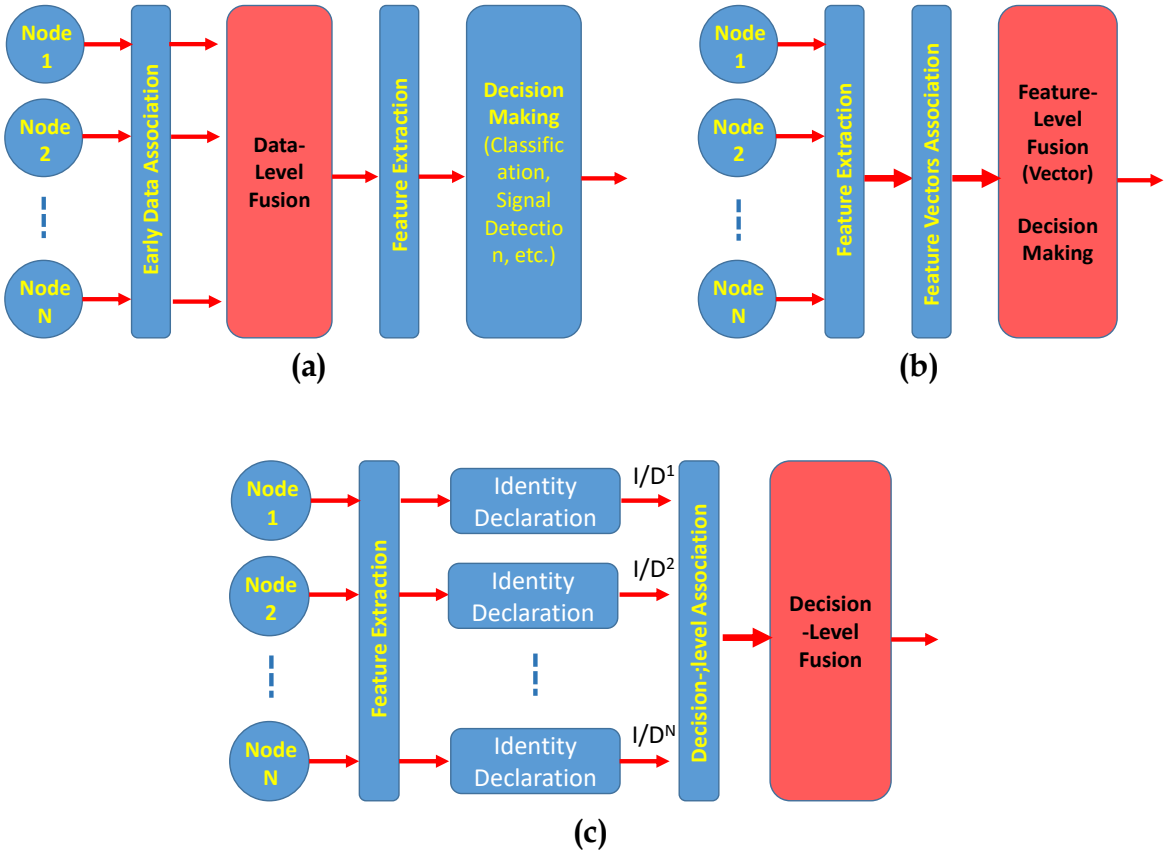


FIGURE 2-1

(a) Raw sensory data fusion (direct fusion). (b) Feature-fusion: combining representative feature vectors extracted from raw sensory data. (c) Decision-level fusion: each sensory data is processed individually to achieve a low-level decision; then accordingly a high-level inference is made. The red box indicates the level of Fusion which is sometimes integrated with Decision process [Hall et al. 2009].

Three General Architectures for Sensor Fusion:

There is no single category for existing Sensor Fusion algorithms. With respect to the level of information that should be combined, and the level of data structure, these algorithms can be categorized into three general architectures:

1. **Early Sensory Data Fusion** (direct fusion): when sensors measure an individual physical property of the environment (e.g., two images taken from single scene to detect depth, or acoustic signals captured by two ears to detect the direction of sound), they can be directly combined into a single decision (FIGURE 2-1 a).
2. **Feature-Fusion:** Sometimes the sensory data cannot directly represent the physical property and needs to be transferred to a feature space. For instance, the direction of sound cannot be directly extracted from the cochlear spikes. Hence, an extra

circuitry is needed to extract Inter-aural Time Difference (ITD[16]) and Inter-aural Level Difference (ILD) to represent direction of sound. ITD and LTD are features that can be later integrated by other direction-related representative features in other modalities, e.g., retinal signals. The features extracted from multiple sensors should be combined into a single concatenated vector (*alignment problem*). This feature vector is the input of the perceptual decision or recognition system (FIGURE 2-1 b).

3. **Decision Fusion:** Fusion can also be made at a high cognitive level. In the decision-level fusion, each individual sensor determines a single preliminary value of an entity's attributes, e.g., location, (referred as identity in FIGURE 2-1), then, these independent values can be fused into a single high-level decision. Fuzzy-Logic and Bayesian Inference Systems, and Voting-based techniques are some examples of the decision-level fusion (FIGURE 2-1 c).

Seemingly, multisensory processing is performed at all the three levels in the nervous system. For instance, retinal information from left and right eyes are combined in early visual cortices to create a rough depth map [Smith and Wall 2008]. Feature-fusion is also widely present in the dorsal pathway [Wand and bend 2012]. High-level decision fusion is also present in the parietal and frontal cortex [Humphreys & Lambon 2015] [Scott et.al 2017]. The theory of Helmholtzian Brain Computing and *Friston's* theory of cortical responses (I adhere to these theories thorough this thesis - see FIGURE 1-5 for more detail) implicitly demonstrate the fact that brain employs sensory integration through different levels of architecture. Where the flow of information is highly interconnected through different cortical areas. However, the mechanisms and the circuitry that implement these processes differ in function and architecture and are just partly understood [Seilheimer et.al 2014].

Over the past two decades, many Sensor Fusion techniques and solutions are developed including: Kalman Filter, Fuzzy-Inference Systems, Neural Networks, Wavelet, Hidden Markov Models, Bayesian Fusion, Voting-based Algorithms, etc. It is hard to categorize these algorithms since many of them are related to each other or differ in terms of generalization, performance, complexity, and flexibility [Sagha et.al 2013]. These developed techniques and mathematical frameworks helped neuroscientist to understand how human brain performs sensor fusion for perception and action generation. Naturally speaking about all the above-mentioned algorithms is beyond the scope of this thesis. Therefore, only those algorithms that are more related to this work

[16] ITD and LTD are two important cues for sound localization which are emerged by head-centered anatomy of ears. ITD is the time difference and ILD is the level difference of sound wave arrived at the left and the right channels, respectively.

and are developed within the scope of Brain Research and Neuro-computing are introduced.

2.2 Deterministic Methods

2.2.1 Voting-based algorithms

The core of voting-based algorithms is the concept of coherency between sensory signals. In fact, coherency is an explanatory quantity that describes the plausibility of inductive perceptual inference [Kiefer 2017]. In other words, when we deal with several sources of information that describe a single physical phenomenon, they must be logically consistent with respect to each other, otherwise there is a defect in a single node and must be compensated and eventually calibrated. In FIGURE 2-2 left, a general scheme of a voting-based algorithm is illustrated. The sensory signals should be evaluated by coherency function (or the voter) to determine which node is inconsistent with respect to the majority of sensory nodes. The term "*majority*" means signals should first agree on a representative value as prototype [Parhami 1996] [Triesch & von der Malsburg 2001] [Desovski et al. 2005]. Then, each single node will be compared with that prototype so that more similarity in attribute indicates better quality and higher contribution in the fusion. As a consequence of this process, the node that is very different or inconsistent with the majority of nodes will be suppressed. Given the quality of each node (described by α_i) and sensory data, sensor fusion algorithm will combine signals. Some algorithms perform this procedure in an adaptive way so that the parameters of the fusion algorithm (e.g., reliability) or coherency function (e.g., prototype) can be changed according to the momentary value of the algorithm outcome [Triesch & von der Malsburg 2001] [Cook et al. 2011] [Axenie & Conradt 2013].

In FIGURE 2-2 right, the mechanism of a voting-based algorithm is demonstrated. In this example, three sensory nodes show three different positions for a single target (S_i). The prototype is the center of mass (average position) as a fair candidate. The Euclidian distance (d_i) of the center of mass and the sensory position of the target are the quantities showing the quality of that sensor. Distance from the center of mass is reversely proportional to the quality of each signal. Identified target position is eventually calculated by a weighted averaging of sensory positions (red dot in FIGURE 2-2 right). Adaptive Democratic Integration that is developed by Triesch & von der Malsburg is one of the earliest voting-based algorithms [Triesch & von der Malsburg 2001]. In the following section, I will describe this algorithm in more detail.

There are cases in which sensory signals do not exclusively reflect an identical attribute and are related to each other with a set of deterministic functions. In this case, the

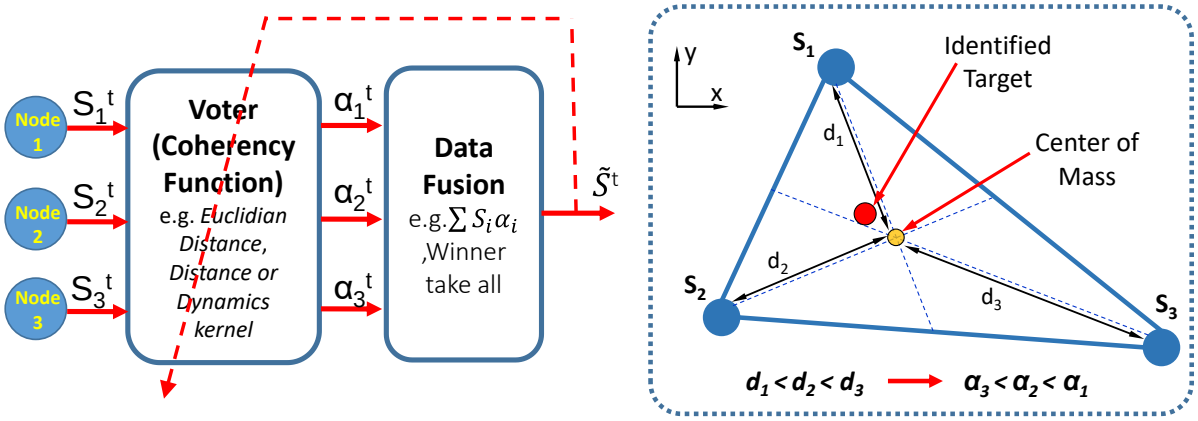


FIGURE 2-2

Left: A general block diagram of a voting-based sensor fusion process. At the first stage and according to a coherency function (e.g., Euclidian distance of each sensory read out from a specific template), the quality of each sensor indicated by α_i will be evaluated. Given α_i, sensory data will be combined. This process can be done in an adaptive way so that the output of the sensor fusion can change the parameters of the coherency function (see Triesch and von der Malsburg 2001).

Right: A simple demonstration of how a general voting-based algorithm works. Three sensors indicate three different positions of a single target. The Euclidian distance (d_i) of each sensory read out and the center of mass (orange dot) is reversely proportional to the quality of each signal. Accordingly, sensory signals will be combined to calculate the identified target position (red dot).

consistency of sensory values can be determined with respect to the relation functions. This function can be learned by the agent while exploring and manipulating the environment. For example, the nervous system automatically knows that the position of an object in retinal coordinate, the position of the eye, and the position of the object with respect to body coordinate, follow a linear relation [Pouget & Sejnowsky 1997]. Or in a manipulation task, the perceived depth of the target object constraints the position and the dynamics of the limbs according to a nonlinear function [Pouget & Snyder 2000]. This special form of voting-based fusion is referred to as Relation Satisfaction in literature. In this chapter, we will see a neural solution for relation learning and relation satisfaction by using Attractor Dynamics. The proposed network is able to learn the determinist relation between encoded cues, and finally perform relation satisfaction [Firouzi et al. 2014a].

2.2.2 Democratic Integration

Democratic Integration is an adaptive algorithm which is developed by Triesch & von der Malsburg in 2001 to identify and track a moving target (e.g., human face) in a video stream. Each cue provides a single attribute of the target (e.g., face color, contrast, motion

velocity, shape) and is associated with an initial prototype. The cues create a set of saliency maps that in fact show the degree of similarity of the image regions to the prototype template (similarity is calculated by using a correlation kernel). In other words, the saliency maps preserve a momentary normalized opinion on the location of the target in image. This opinion is associated with a momentary reliability weight by which a combined estimate of the target location can be calculated. The initial combined estimate is a weighted average of individual saliency maps using reliability weights. Eventually, winner-take-all (maximum probability) applied on a combined estimate map indicates the location of the target in the current image.

Given the identified target location, the reliability of each cue and its associated prototype window will be updated for the next images. This process enables an adaptive change of each representative attributes, so that, in case the scene properties change abruptly in time (e.g., a sudden change in light condition, subject turning, one sensory node drop-off) the system can successfully recover the target position using other cues and calibrate sensor prototypes simultaneously. The quality of each cue, which defines how well that cue predicts the target successfully, is calculated by subtracting the cue's saliency map value at the target position (winner location) and mean saliency value for that cue. Then, the momentary reliability value of that cue is updated toward the calculated quality factor, with a specific time constant. The time constant should be large enough to make the system robust against noise, and small enough to allow for quick adaptation. Similarly, to calibrate the discordant cue, the prototypes also follow a dynamic toward the attribute values (prototype window) around the winner location (current identified target). To prevent one cue to take-over the whole perception process, a carefully defined quality function is proposed.

Democratic Integration provides a powerful fault-tolerant framework for feature-level image fusion. However, it is necessary to predefine a set of prototypes for each cue, while the weighting mechanism does not directly reflect the governing noise process. In fact, the quality of each signal is not defined as mathematically reasonable as the reliability in MLE or Kalman filter algorithms (see Section 2.3.4). The main advantageous characteristic of this algorithm is the adaptive calibration mechanism of the sensory nodes with respect to the winner location attributes. In Section 4.2.5, the results of MLE with a voting-based algorithm is compared and analyzed in a heading estimation experiment.

2.2.3 Relation Satisfaction

A key requirement for any systems including biological or man-made systems, is to interact properly with their environment, and to estimate physical properties of the real world through partially reliable observations. For instance, to reach an object by hand, one must configure the arm joints with respect to the visual location of the object and

proprioceptive cues. Apart from intrinsic variability of neural activity and sensory data, accessible sensory cues are often partially observable. The human brain can combine these partially reliable and partially observable pieces of information to optimally estimate the state of the world and consequently handle motor tasks efficiently [Ernst and Bülthoff 2004] [Simoncelli 2009]. The general architecture of voting-based algorithms is feed-forward (FIGURE 2-2 left). However, feedforward processing is not the only canonical form of information processing in the brain [Miller 2016]. The brain is composed of a highly distributed and interconnected architecture, where the feedforward stream of sensory information is usually modulated by feedback connections [Miller 2016]. As we discussed in Chapter 1, this recurrent connection enables a powerful computational mechanism in the cortex to retrieve partially observable information (hidden states and concepts in FIGURE 1-5). When a weak or a noisy sensory input drives the dynamics of the recurrent circuit, it results in a fast interpretation of the world before sensory feedback provided (see predictive perception discussed in Section 1.2.2).

On the other hand, multimodal sensory data are mostly different in terms of physical properties, dynamic range, and are initially presented on separate coordinates. Therefore, they must be aligned into a common frame of reference. This coordinate transformation is essential for spatial perception and can be formulated by a set of deterministic relations. This relation function constraints and governs the coherency and consistency between sensory values and follows a dynamic process called relation satisfaction. Relation satisfaction means all momentary sensory experiences that are mostly noisy must satisfy the rational relations to compensate possible uncertainties, inconsistencies, and sensory defects or deprivations. Relations amongst sensory cues can be discovered and learned by performing action-perception loop. Consequently, if one sensory node drops off, the other sources of information can restore that deprived information given the governing relations and recurrent exchange of information. One famous example is the linear transformation of retinal frame of reference to the head-centered frame of reference, modulated by eye motion (see also FIGURE 1-7), [Pouget and Sejnowsky 1997] [Brosteck et al. 2015].

2.2.3.1 Interacting-Maps Network for Fast Visual Interpretation:

Cook et al. demonstrate the viability of a computational approach for fast visual interpretation by using relation satisfaction principle [Cook et al. 2011]. In FIGUER 2-3 the instantiation of this approach is illustrated, where a 128*128 neuromorphic silicon retina provides the network with dynamic features of the scene. Dynamic features include any changes, either negative or positive changes, in light intensity. Each single cell of the retina generates a single spike (or event) whenever a relative change in light intensity exceeds a pre-set threshold (positive event is associated with positive change and negative event reflects a negative change in the light intensity). V in FIGURE 2-3 which is

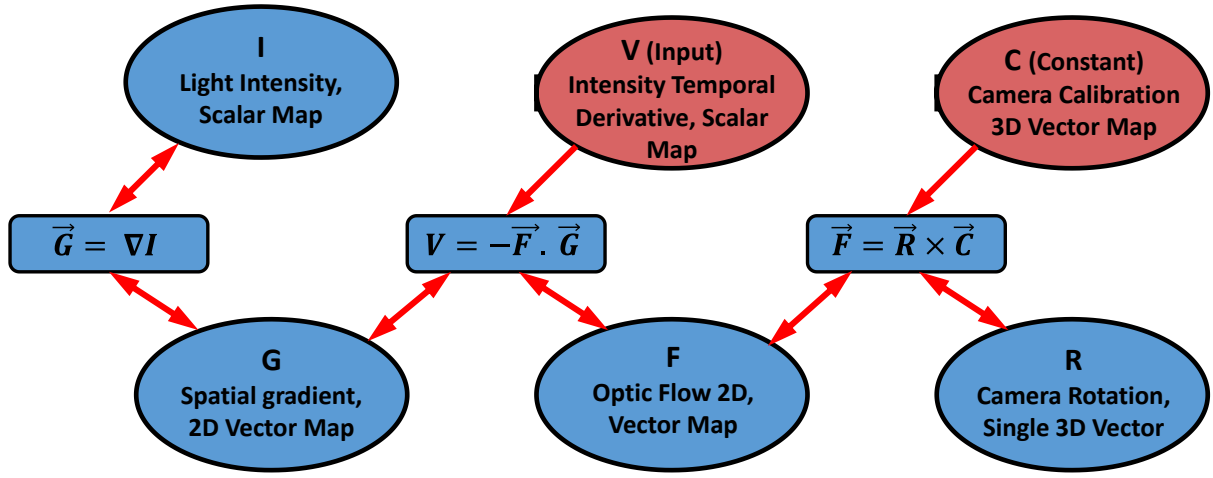


FIGURE 2-3

A distributed relation-satisfaction network for fast visual interpretation proposed by Cook et al. [Cook et al. 2011]. The relationships between internal and sensory variables are shown by rectangles and are applied for each single pixel independently. The picture is borrowed from [Cook et al. 2011] with permission and minor changes.

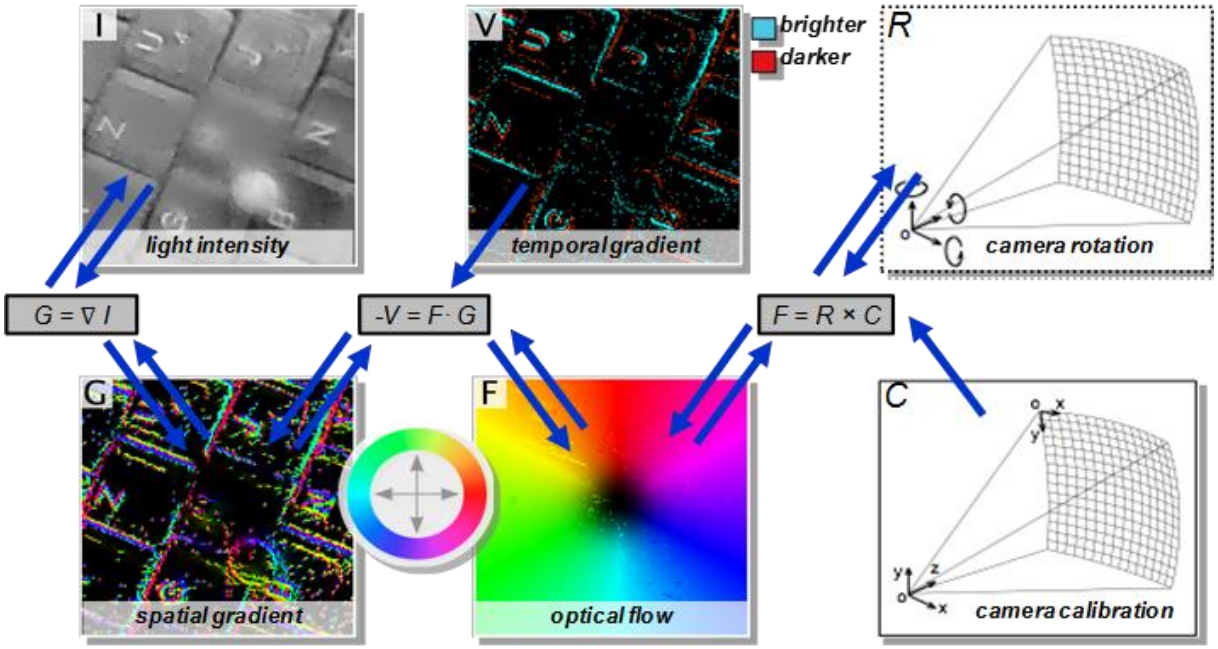


FIGURE 2-4

A sample result of the network behavior after reaching stable state. *V* shows the input; *I* presents internally estimated intensity; *G* shows the estimated spatial gradient. Pattern of flow motion, F is also color-coded. Taken from [Cook et al. 2011] with permission and minor changes.

the temporal intensity derivative ($\frac{\partial I}{\partial t}$), is the sole input to the network taken from silicon retina. *C* is camera calibration map which is assumed constant, and *R* is a single three-dimensional vector shows an estimate of camera rotation. *F* is flow motion or optic flow vector map ($\frac{\partial x}{\partial t}, \frac{\partial y}{\partial t}$), and *G* is intensity spatial gradient ($\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}$). The intermediate maps (blue ellipses in FIGURE 2-4) are initialized randomly and the update rules are derived based on relations. To update a single map given the connected relations, it is assumed that the connected maps provide correct information. Given the connected map values, a candidate map is computed that should satisfy the relation as much as possible. Then, the map value is updated by taking one small step toward the calculated candidate map. In case the reverse of relations cannot be directly calculated, the step of update rule is determined by the derivative of a quadratic error function. For instance, *I* and *G* are related according to the following relation:

$$G_{x,y} = \nabla I = \begin{bmatrix} I_{x+1,y} - I_{x,y} \\ I_{x,y+1} - I_{x,y} \end{bmatrix} \tag{2.1}$$

Therefore, the update rule for *G* given the intensity map *I* is simply defined by equation 2.2, where $0< \delta_{IG} <1$ indicates the small update factor:

$$G_{x,y}^{t+1} = (1 - \delta_{IG}) G_{x,y}^{t} + \delta_{IG} \nabla I_{x,y}^{t} \tag{2.2}$$

To update the intensity map *I*, given spatial gradient map *G*, we define first an error map *E* and then, the derivative of error ΔE:

$$\mathrm{E}_{x,y} = G_{x,y}^{t} - \nabla I_{x,y}^{t} \tag{2.3}$$

$$\Delta \mathrm{E}_{x,y}^{(x)} = \mathrm{E}_{x,y}^{(x)} - \mathrm{E}_{x-1,y}^{(x)}, \quad \Delta \mathrm{E}_{x,y}^{(y)} = \mathrm{E}_{x,y}^{(y)} - \mathrm{E}_{x,y-1}^{(y)} \tag{2.4}$$

Finally, the intensity map *I* is updated according to the estimated derivative of error:

$$I_{x,y}^{t+1} = (1 - \delta_{GI}) I_{x,y}^{t} + \delta_{GI} (I_{x,y}^{t} - \Delta \mathrm{E}_{x,y}^{(x)} - \Delta \mathrm{E}_{x,y}^{(y)}) \tag{2.5}$$

As shown in FIGURE 2-4, most of the internal variables are not directly observable but the relation satisfaction dynamics help to restore grayscale intensity map and optic flow. This problem is not a trivial problem to solve. However, the purpose of this network is not to calculate the intensity map and flow-motion that can be solved by sophisticated methods. The main goal of this network is to investigate the viability of the interacting architecture for relation satisfaction and creating a coherent interpretation of the world. Furthermore, it is also questioned how possibly interacting cortical areas can facilitate this computation for perception. The architecture of this network resembles deep belief network that formalizes Helmholtzian brain computing theory (see Chapter 1). The major difference is that, the background internal beliefs in interacting-maps network is

represented by relation functions rather than probability functions. This form of multisensory inference enables a fast-explanatory interpretation of the world, by which the nervous system does not need to wait for sensory inputs to be well-digested in poly-sensory areas and then, to initiate the action. In fact, a weak sensory evidence can activate a distributed hierarchical circuit, and thereby, internal hidden variables, so as consequently to create a coherent internal belief in the state of the world. This idea is the central hypothesis of this thesis which is widely supported by Neurophysiological evidences. Musacchia et al. showed a persistent neural activity in brainstem just 15ms after stimulus onset during multimodal speech perception [Musacchia et al. 2005]. Moreover, it is evident that salient sensory events can directly reach multimodal cortical areas by bypassing early and secondary sensory areas [Liang et al. 2013]. This fast information transmission is facilitated by a direct thalamo-cortical pathway to higher cortical areas, parallel to the pathway through primary sensory cortices. Although this pathway carries salient information that demands a fast reaction, it challenges the primitive notions of multisensory integration as a procedural process in the brain; where the stimuli should follow a hierarchy to reach the association cortex [Liang et al. 2013] [Paraskevopoulos and Herholz 2013].

Even though this network shows promising results, Cook et al. left the question open how possibly a plausible neural circuit can perform this form of computation. Moreover, it is also challenging to represent and store arbitrary relations through synaptic weights, especially in the presence of higher order modalities [Cook et al. 2011]. Cook proposed an unsupervised learning framework for relation learning between two interacting maps, each includes a population of neurons to encode a single cue [Cook et al. 2010]. The neurons of each population are laterally connected, and they are also mutually connected to another population. This simple network is able to perform relation learning and relation satisfaction for two encoded variables. But, it is not clearly discussed whether this scheme is scalable for higher order relations? In Section 2-3 a flexible framework for relation learning is proposed using attractor dynamics. The evaluation results demonstrate the feasibility of this approach in a neutrally plausible way.

2.2.3.2 Cortically Inspired Sensor Fusion Network for Heading-Estimation:

Another issue in multi-sensory research which is less investigated is how to encode and learn the reliability of the sensory signals in neural models. Sensory cues do not exhibit identical reliabilities and they might change in time. Axenie and Conradt have proposed a distributed network of sensory nodes, synonymous to Cook's integrating-maps network, to investigate how possibly the reliability and sensory defects can be automatically detected [Axenie and Conradt 2013]. The basic idea is borrowed from the interacting-maps network, where sensory nodes are modeled as a group of representing

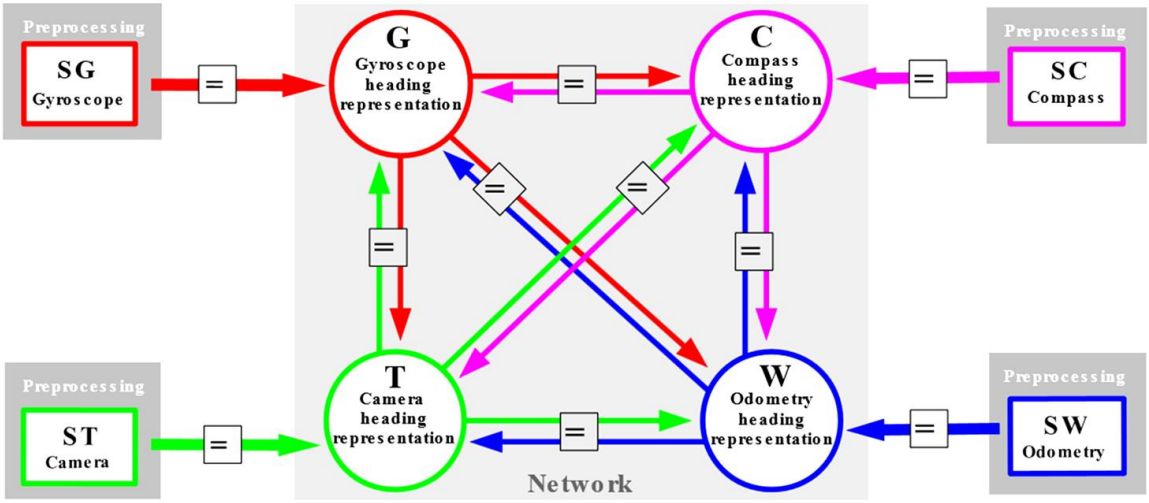


FIGURE 2-5

The basic architecture of Cortically-Inspired-Sensor-Fusion network proposed by [Axenie and Conradt 2013]. The network is about to estimate heading angle in an omni-direction mobile robot which is equipped by four sensors (Gyroscope, Compass, Odometery, and Vision). Since sensors are not directly measuring the heading angle, a pre-processing stage calculates the sensory-specific estimate of heading.

maps and are connected to each other according to a specific deterministic relation (FIGURE 2-5). The reliability of cues is also dynamically updated similar to the mechanism applied in the democratic integration algorithm [Triesch & Von derMalsburg 2001]. The difference is the replacement of the prototypes by relations. In FIGURE 2-5, the basic architecture of this approach is depicted, in which, four mutually connected sensory nodes (circles) represent independent estimates of a mobile robot heading. Since sensors are not directly measuring the heading angle, a pre-processing stage (rectangles) calculates and delivers the sensory-specific estimates of the heading. Moreover, the sensory nodes are about to measure an identical property (heading angle), so the governing relation function is equality. This problem is sometimes referred as sensory convergence in the literature, where multiple sensory modalities are integrated to measure a single variable [Hartline 1988].

Axenie and Conradt developed a gradient-decent based rule to update sensory reliability [Axenie and Conradt 2015]. According to this rule, first the mismatch between the current sensory value and the agreed value (after relaxation) is computed, and then, the reliability is updated toward a normalized factor which is reversely proportional to the calculated error. On the other hand, the update rate (δ in equation 2-2) is proportional to the reliability. That means the relaxation dynamics is in favor of more reliable node.

This dynamic converges to a value that indicates how much the respective node is in agreement with the other nodes according to the governing relations. The main advantage of this approach is its simplicity, flexibility and robusticity against drop off. However, the sensitivity of the updating rule to the derivative of relation function makes it infeasible for highly nonlinear relations, where the dynamics of the network becomes unpredictable. On the other hand, the defined reliability is not able to formulate the statistical fluctuation of the sensory nodes. Consequently, there might exist a noisy node that satisfies the relation functions, so thereby takes over the final estimate, and contaminates the estimate by noise. In Section 4.2.5 this issue is shown by comparing the outcome of this network with the Maximum Likelihood Estimation [Firouzi et.al 2014a].

2.2.3.3 Mutual Prediction

In general, the real-world variable is hidden behind the sensory read out: either that could be corrupted by noise or partially observable. The nervous system thus has no direct access to any teaching signals to identify how much of the signal is noise or bias and how much information content is carried. When there are multiple sources of information that captures different aspects of the real-world but are highly correlated, then, there is a chance for the nervous system to predict one sensory information from sensory sources. This property is referred as *mutual predictivity* in the literature [Berger 2006]. *Mutual productivity* of two independent sources S_1 and S_2 (each measure two different attributes of a real-world entity) means that there are two hypothetical functions f and g such that $f(S_1)$ is correlated with S_2 and g (S_2) is correlated with S_1. Now based on these correlation functions, each sensory node can predict the value of other sensory nodes. As a result, combining the real sensory observation with predicted value enhances signal quality. This technique is called *Sensor Fusion by Mutual Prediction* or *model fusion*. From another point of view, mutual prediction can also be a specific relation satisfaction problem in which the governing relations are determined by the correlation functions f and g.

Berger argues that, if two interconnected cortical columns preserve information from different senses, the mutual interconnection of layer 2/3 can possibly implement mutual prediction such that each sense can mutually predict and test the value of the other modality [Berger 2006]. It is also evident that this mechanism can be possibly employed by some parietal neurons to register the tactile and the respective visual cues into a single frame of reference [Avillac et al. 2005] [Wright & Glasauer 2006].

2.2.4 Image Fusion

Image fusion is one of the most famous hallmarks of sensor fusion in engineering. For instance, combining multiple spectral images into a single image has been widely used for signal detection. Or combining multiple images from a single scene that helps to

extract the depth information. One specific example of image fusion is stereoscopic fusion in which a pair of images taken from different angels are used to solve the correspondence problem (see Section 3.1.1) and thereafter extract the 3D information out of that [Hartly & Zisserman 2003]. Most of the algorithms in classical vision focused on frame-based image fusion that utilizes the relative disparity to estimate depth. It is still not well understood how cortical circuit computes retinal disparity. On the other hand, neural fibers from retia to thalamus and cortex, send visual information asynchronously and not frame-wise. This is a new challenge that calls for sophisticated event-based algorithms that might help to understand the functionality of human visual system and enable new technologies. In Chapter 3, using the principle of cooperative computing, an event-based neural model is proposed in order to solve the problem of stereoscopic fusion. We have used a silicon-retina sensor that imitates the way human retina encodes visual information.

2.3 Probabilistic Algorithms for Sensor Fusion

To interact with real-world, the perceptual system should cope with unavoidable obstacles like *imprecision*, *inaccuracy*, and the *not-directly-observable* state of the world (hidden real-world variables). Although the underlying models of perception are context-dependent, one of the key factors that should be considered in a model is the intrinsic stochasticity of the real world. The intrinsic inaccuracy and imprecise nature of sensory world, and motor system can be well accommodated within probability theory. It is discussed in Section 1.1.2 whether the existing stochasticity in neural activities is functionally related to the probabilistic world. What would happen if we have a deterministic brain? And if it was the case, how difficult that will be to deal with our simple tasks. A wide range of human and animal cognitive functions including Multisensory Perception, is well modeled by probabilistic frameworks mainly anchored in the Bayesian Decision Theory. This theory relates the belief of possible existing hypothesis to current sensory or motor evidences (observations) and the initial or prior information. In conjunction with Helmholtzian theory of brain and Friston theory of cortical computing, the nervous system is a pool of conditional beliefs which models and controls our world (see FIGURE 1-5), and that can be best formalized particularly within Bayesian framework [Kiefer 2017]. In this section, we will summarize the basic probabilistic algorithms for the Multisensory Integration and a taxonomy of Bayesian Perception in different contexts.

2.3.1 Maximum Likelihood Estimation

One way to describe the belief of existing hypothesis H^S, given n sensory observations $\{S_k \mid k=1, 2, ..., n\}$, is to compute the probability[17] of the sensory evidences, if the hypothesis H^S is present. This function is called the likelihood function and can be described by the following equation:

$$L(H^S) = P(S_1, S_n, \dots, S_n \mid H^S) \tag{2-6}$$

Since the underlying neural processing of each modality is mostly independent, the model of noise for each source of information is assumed to be independent [Burge et.al 2008] [Ernst & Di Luca 2011]. So, Eq. (2-6) can be re-written as bellow:

$$L(H^S) = \prod_{k=1}^{n} P(S_k \mid H^S) \tag{2-7}$$

Obviously, the sensory measurements are known (unlike the real state of the world), so the likelihood function can be parametrized by the governing noise process of each sensory node. So, the best possible[18] hypothesis of the current state variable (or hidden state) for the set of current observations S_k is the one that maximizes the likelihood function or equivalently the root of partial derivative of H^S:

$$\hat{S}^{MLE} = \{S \mid \frac{\delta L(S)}{\delta S} \approx 0\} \rightarrow \prod_{k=1}^{n} \frac{\delta P(S_k \mid S)}{\delta S} = 0 \tag{2-8}$$

For any known (or modeled) arbitrary noise process, the root of (2-8) determines the MLE estimate of the state variable S, given the fact that n-sensory nodes are measuring and identical physical variable within different attributes that are statistically independent. Now let us assume that k-*th* sensory node has an additive noise with normal distribution: $N(b_k, \sigma_k)$. So, the likelihood function will be as below (since a single hypothesis H^S is equal to a single possible hidden state S, H in (2-7) is replaced with S):

$$L(S) = \frac{1}{(2\pi)^{n/2} \prod_k \sigma_k} \prod_{k=1}^{n} e^{-\frac{(S - S_k + b_k)^2}{2\sigma_k^2}} \tag{2-9}$$

A good way to simplify equation (2-9), is to transform the product-of-exponential factors into sum-of-quadratic terms using logarithm function. On the other hand, the root of likelihood's derivative is equal to the root of log-likelihood's derivative; because of two reasons: first, $L(S)$ is positive infinite, and second, log is a monotonically decreasing transformation. $LL(S)$ in (2-10) is the log-likelihood function of (2-9). Note that, for the simplicity, S_k - b_k is replaced with μ_k:

[17] Usually the likelihood probability function is not normalized to sum up to 1, unlike posterior or prior probability function.

[18] The "best hypothesis" here denotes the most probable one that describes the current state of the world for us.

$$LL(S) = -\frac{n}{2}\log 2\pi - \sum_k \log \sigma_k - \sum_k \frac{(S-\mu_k)^2}{2\sigma_k^2} \; where \; \mu_k = S_k - b_k \tag{2-10}$$

Given (2-8) and (2-10), the derivative of *LL(S)* and the MLE estimate will be determined by the following equations:

$$\frac{\delta LL(S)}{\delta S} = -2\sum_k \frac{(S-\mu_k)}{2\sigma_k^2} = -[S\sum_k \frac{1}{\sigma_k^2} - \sum_k \frac{\mu_k}{\sigma_k^2}] \tag{2-11}$$

$$\frac{\delta LL(S)}{\delta S} = 0 \;\rightarrow\; \hat{S}^{MLE} = \sum_{k=1}^{n} w_k \mu_k \,, \; where \; w_k = \frac{\frac{1}{\sigma_k^2}}{\sum_k \frac{1}{\sigma_k^2}} \; and \; \sum_{k=1}^{n} w_k = 1 \tag{2-12}$$

As shown in (2-12), the MLE estimate $\hat{S}^{MLE}$ is the weighted average of the mean values for each individual sensory node. The weight for a single node is reversely proportional to its respective variance σ_k^2. These weights, w_k in (2-12), are in fact the respective reliabilities of each sensory signal or the degree of its contribution in the combined estimate - $\hat{S}^{MLE}$. Since MLE estimate is the sum-of-product of n Gaussian random variables $S_k \sim N(b_k, \sigma_k)$, it can be seen as a Gaussian random variable with a mean value equal to $\hat{S}^{MLE}$. The variance of $\hat{S}^{MLE}$ is equal to:

$$\sigma_{MLE}^2 = \frac{\prod_{k=1}^{n} \sigma_k^2}{\sum_{k=1}^{n} \sigma_k^2} \; or \; \frac{1}{\sigma_{MLE}^2} = \sum_k \frac{1}{\sigma_k^2} \tag{2-13}$$

The variance of MLE estimate shown in equation (2-13) is smaller than the variance of each individual sensory signal. Because the inverse of variance for MLE estimate is equal to the sum of inverse-of-variance for each single sensory node. This reflects the main benefit of MLE fusion algorithm. In FIGURE 1-6, it is demonstrated why MLE is an optimal variance-minimizing strategy for sensor fusion; where the variance of Likelihood functions of final estimate for two different weighting scenarios are compared: one for MLE weighting scheme (equation 2-12) and the other for a down-weighted visual signal (black and blue curves respectively).

Now if we expand μ_k in (2-12), we would have two terms for MLE estimate: the first term is the reliability-based weighted sum of sensory signals, and the second term is the weighted average of the bias b_k for the associated sensory nodes. Bias is often deterministic and constant[19] during each trial. So (2-12) can be re-written as follows:

$$\hat{S}^{MLE} = \sum_{k=1}^{n} w_k S_k - \sum_{k=1}^{n} w_k b_k \tag{2-14}$$

The second term in (2-14) shows the main disadvantage of MLE. If one sensory node with high value of statistical reliability contains a big value of bias, then, the final estimate will be drastically drifted far away from real-world signal. So, this algorithm is optimal just

[19] Assuming that the sensory noise is a "stationary process" in which the parameters of the process including variance and mean are changing trial by trial.

under special circumstances in which the sensory nodes are unbiased and are statistically independent. There are several psychophysical experiments reported how human perceptual system employs MLE to integrate information across senses or within a single modality to achieve a statistically optimal estimate of the world attributes (e.g., visual-haptic size estimation [Ernst and Banks 2002], visual- acoustic localization [Alias and Burr 2004], and retinal-disparity and motion-cues integration for depth estimation [Bradshaw and Rogers 1996]).

MLE in general can be used as a hypothesis testing framework, and sometime is used for parameter identification of a sensory system [Myung 2003]. But, for that we need to know the process of noise, e.g., Gaussian.

2.3.2 Basic Bayesian Integration

The world natural attributes and features within different senses follow a structural regularity. For instance, our visual system is stimulated by more horizontal and vertical edges every day than any other intermediate angles [Girshick et.al]. That means the distribution of sensory stimulation encountered in the real-world is often non-uniform. More interestingly, it is evident that early visual cortex recruits more neuros and resources to code horizontal and vertical edges [Sadeh and Rotter 2014]. This fact supports the notion that says: *"the statistics of the world must be internalized and encoded within the nervous system"* [Simoncelli 2009].

The prior knowledge about the sensory world is evident before facing with any sensory evidences, and that should be incorporate into our perceptual system. Bayes rule can formalize this process and relate the probability of a real-world variable to the current sensory evidences of that variable and the prior information about that variable (frequency of the stimulus). This probability is called posterior probability and is described in equation (2-15), where *P(S)* is the prior probability and the sensory evidence is described by the likelihood function similar to (2-6):

$$P(S|S_1, S_n, \dots, S_n) = \frac{P(S_1, S_n, \dots, S_n|S)P(S)}{\int P(S_1, S_n, \dots, S_n|S)P(S)dS} \tag{2-15}$$

The integral term in the denominator is a marginalization process over S variable to compute the joint probability of the current sensory evidences S_k. This term is a normalization value and can be neglected. So, a non-normalized posterior can be described as:

$$P(S|S_1, S_n, \dots, S_n) \propto P(S_1, S_n, \dots, S_n|S)\, P(S) \tag{2-16}$$

Similar to the analysis described in Section 2.3.1 and with the same assumptions for sensory nodes (statistically independent and normally distributed), posterior probability function will be as bellow:

$$P(S|S_1, S_n, \dots, S_n) \propto \prod P(S_k|S)\, P(S) \tag{2-17}$$

If we assume the noise process in each sensory stimulus is not uniform and is normally distributed (or equivalently the prior probability is a Gaussian function), the posterior probability function will be also Gaussian, since the production of multiple Gaussian functions is also a Gaussian (μ_p and σ_p are the mean and standard deviation of the prior distribution respectively):

$$P(S|S_1, S_n, \dots, S_n) \propto \frac{1}{\sqrt{2\pi}\sigma_p (2\pi)^{n/2} \prod_k \sigma_k} e^{-\frac{(S-\mu_p)^2}{2\sigma_p^2}} \prod_{k=1}^{n} e^{-\frac{(S-S_k+b_k)^2}{2\sigma_k^2}} \tag{2-18}$$

Having the prior knowledge incorporated, and similar to MLE estimate, we can define an estimate of the real-world variable S that maximizes posterior probability instead. This fusion technique is called Maximum-A-Posterior estimation or MAP:

$$\hat{S}^{MAP} = \sum_{k=1}^{n} w_k \mu_k + w_p \mu_p, \ where\ w_k = \frac{\frac{1}{\sigma_k^2}}{\sum_k \frac{1}{\sigma_k^2} + \frac{1}{\sigma_p^2}}\ and\ w_p = \frac{\frac{1}{\sigma_p^2}}{\sum_k \frac{1}{\sigma_k^2} + \frac{1}{\sigma_p^2}} \tag{2-19}$$

$$\frac{1}{\sigma_{MAP}^2} = \sum_k \frac{1}{\sigma_k^2} + \frac{1}{\sigma_p^2} \tag{2-20}$$

As shown in (2-20), introducing the prior knowledge into the MLE will result in an enhanced reliability, because the variance of $\hat{S}^{MAP}$ is smaller than $\hat{S}^{MLE}$. However, the final estimate will be drifted towards a-priori[20] expectation, μ_p[21]. This phenomenon, that is referred as perceptual illusion (or bias), is empirically reported in several studies [Kersten et al. 2004] [Stocker & Simoncelli 2006] [Körding & Beierholm 2006] [Rohe & Noppeney 2015]. In FIGURE 2-6, the results of a Monte-Carlo simulation for *MLE* and *MAP* fusion of two attributes of a single stimulus (Visual and Acoustic location) is illustrated. As shown in this figure, the variance of *MAP* is reduced compared with MLE, but at the cost of perceptual bias toward the prior expectation value S_{pri}. On the hand, it is inevitable for both algorithms to avoid the destructive effect of sensory bias in the final estimate. In general, this basic form of Bayesian Integration is still error prone. So, this is one of the costs of integration that must be balanced with the benefit of variance minimization [Ernst & Di Luca 2011].

Prior is a very important entity for perceptual system and usually reflects the occurrence-frequency of the sensory and sensory-motor experiences. In a real-world behavior, we basically intend not to change our prior expectations rapidly. In fact, our perception of the world is highly subjective and biased towards our subjective expectations that our nervous system learnt. The neural correlates of the prior-induced bias are questioned. It appears that the prior expectation should be registered within the

[20] Reasoning or knowledge which proceeds from theoretical deduction rather than from observation or experience.
[21] This form of bias is sometimes referred as perceptual illusion in the literature.

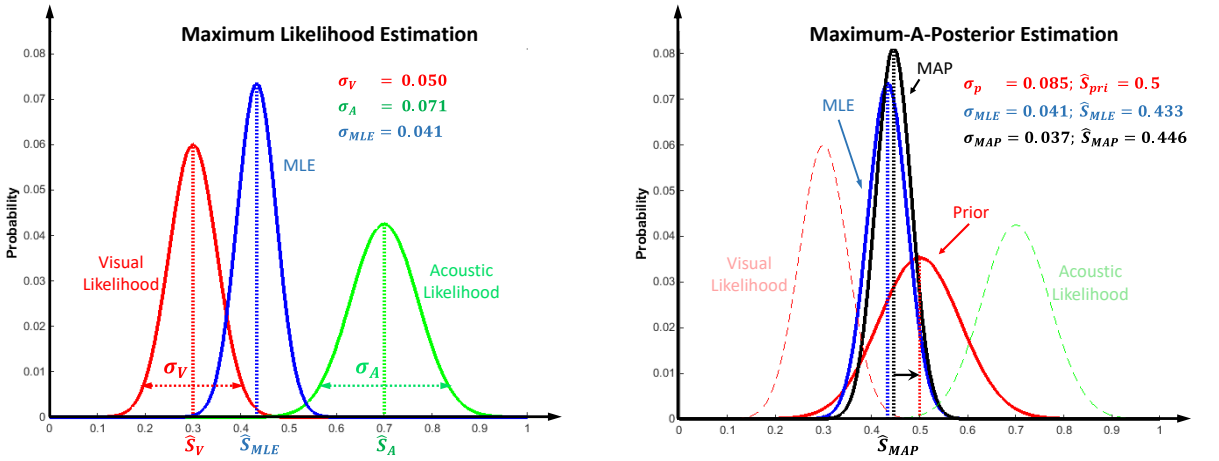


FIGURE 2-6

A comparison between MLE estimation and MAP estimation and incorporation of the prior in fusion. **Left:** a representation of MLE estimation given two independent sensory information corrupted with additive Gaussian noise. **Right:** incorporating the prior into MLE fusion, which results in MAP estimation. The black Gaussian profile represents the distribution of MAP estimate and demonstrates the reduced variability of MAP compared with MLE (blue Gaussian profile). The perceptual bias in MAP estimate ($\tilde{S}_{\text{MAP}}$) is also shown by a drift towards the prior (red Gaussian profile). The black arrow represents the direction of the prior-induced perceptual shift. Both graphs are generated through a Monte-Carlo simulation with 10^6 samples, and the likelihood functions are normalized.

hierarchical structure of our perceptual system, from high level concepts to low level features [Bubic et.al 2010 In Section 2.3.1and 2.3.2, we show]. For instance, the number of recruited neurons for edge detection in ventral pathway follows the statistics of line-features which exist in a wide range of natural images [Girshick et.al. 2011]. However, it is still not fully understood where in the sensory cortices and through which neural mechanism, prior information is preserved and learnt.

2.3.3 Integration breakdown and Recalibration using Coupling-Prior Model

In the process of perception - whether multisensory or uni-sensory - CNS always must incorporate two components, likely according to the Bayes rule:

- **Prior**, that remains plastic and reflects the statistics of a cross-modal or a within-modal sensory stimuli in the environment.
- **Sensory likelihood** or sensory evidence, which is provided by the momentary and partially reliable sensory observations.

In Section 2.3.1 and 2.3.2, we show how the basic Bayesian Integration can systematically combine the prior and the sensory evidence to compute a single estimate of the real-world

state variable and/or internal state of the body. The main benefit of the Bayesian integration is to minimize the noise-driven imprecision in the final estimate, as provide in equation (2-20), in which σ_{MAP} is smaller than the variance of each sensory estimates[22] σ_A and σ_v. However, if one of the sensory nodes exhibits a drift with respect to the real state variable S, then, the Bayesian *MAP* and *MLE* fusion will inevitably introduce that sensory bias[23] in the final estimate (see equation 2-14 and 2-19). Therefore, *MAP* and *MLE* are no longer optimal and must be compensated properly. In fact, minimizing the bias-induced inaccuracy and noise-driven imprecision are two competing factors within the basic *MAP* and *MLE* fusion (see FIGURE 2-6). So, the basic Bayesian integration which is described in Section 2.3.2, should be modified to employ a mechanism to reduce the cost of the integration as optimal as possible. The source of bias, modeled by b_k in equation (2-18), is potentially due to

(i) external factors such as the effect of humidity in the speed of sound, glass-induced light refraction, the effect of sub-zero temperature or wearing-gloves in sense of haptic.

(ii) (ii) internal influences such as muscle fatigue, temporary sensory deprivation, deficits. Given the current sensory evidences, the *reliability assignment problem* in equation (2-14) and (2-19) is not a difficult task for CNS, because the noise content is present within the sensory information and can be simply measured online by the nervous system. Whereas the real value of the stimulus is hidden, and the sensory signals also do not carry any direct information about their inaccuracy. As a result of that, the systematic bias cannot be directly measured from sensory likelihood and even the final estimate $\tilde{S}^{MAP}$. Even the discrepancy between a pair of sensory estimates S_i and S_j cannot determine which node is inaccurate. Because that is also a random variable and it is changing from one trial to another.

However, discrepancy is a useful cue for the perceptual system to utilize a computational strategy and to avoid inaccuracy in the final estimate as optimal as possible. Larger the discrepancy or sometimes referred as sensory conflict becomes, it is more rational[24] and optimal to stop fusion over sensory measurements S_i, hence, to break down the integration. But it is still tricky to determine a quantitative threshold of the sensory conflict for the integration *breakdown*. This problem is called C*redit-Assignment* in sensor fusion, through which the perceptual system needs to determine the reason of current sensory conflict, and then, accordingly should perform fusion or break it down. The process of fusion-breakdown is called *Segregation*, opposite in meaning to Fusion. If the sensory conflict stays persistent across the trials, then, it is more likely due to

[22] The preliminary sensory estimate, given the sensory evidence is MLE.

[23] Here we are considering only the additive bias.

[24] Rationality is defined as coherency-maximization in perception [Kiefer 2017]

inaccuracy and must be compensated within Segregation and later Recalibration. And if the conflict fluctuates randomly, it is likely caused by sensory noise that can be mitigated by optimal algorithms like linear fusion, i.e., *MAP*.

The belief in whether the noise or bias is the reason of discrepancy, no matter how small it is, can be best expressed within a probabilistic framework. To drop this notion into a computational model, Ernst and Di Luca introduced an extended Bayesian framework that employs a strategy to balance the cost and benefit of multisensory fusion [Ernst and Di Luca 2011]. In this model, a statistical mapping between sensory estimates S_i is defined that reflects the jointly encountering distribution of the signals, and also implicitly the belief in whether the sensory discrepancy is due to noise or bias. This mapping is called *coupling-prior* and is used as a prior distribution in the Bayesian Integration. At the following sections, we will explain how this model deals with (i) the credit-assignment problem, (ii) the balance of cost and benefit of integration, (iii) the integration breakdown in case of large sensory conflicts, and finally (iv) the calibration of persistently biased sensory nodes. This model is one of the first mathematical frameworks that has integrated three fundamental functions of multisensory integration - i.e. optimal fusion, integration breakdown, bias estimation, and calibration - within a unified model [Ernst 2005] [Ernst 2007] [Bug et.al 2008] [Ernst and Di Luca 2011].

2.3.3.1 A unified Model for Fusion, Partial Fusion, and Segregation

Similar to the size discrimination task described in [Ernst & Di Luca 2011], we assume two physical attributes, Visual S_w^V, and Haptic S_w^H are captured by sensory system to measure the size of an object S_w. Given the physical attributes $S_w = (S_w^V, S_w^H)$[25], let $S = (S_V, S_H)$ be the sensory signals that are biased with respect to S_w, i.e. $S = (S_w^V + B_V, S_w^H + B_H)$. To represent the sensory evidence in a 2-Dimentional space, and assuming an additive noise process with normal distribution for each modality, the joint likelihood is defined as:

$$P(z_V, z_H | S_V, S_H) = N(S, \Sigma), \Sigma = \begin{bmatrix} \sigma_V^2 & 0 \\ 0 & \sigma_H^2 \end{bmatrix} \tag{2-22}$$

Where, N is a bivariate Normal distribution with covariance matrix Σ, and (z_V, z_H) is the current sensory measurement that is infected by noise. As a result of an additive Gaussian noise process, the mean of likelihood function is the MLE estimate: $\hat{S}^{MLE} = (\hat{S}_V, \hat{S}_H) = (z_V, z_H)$. The Gaussian bump in the left column of FIGURE 2-7 represents a hypothetical likelihood function, in which the MLE estimate is indicated by a black cross. it is important to note that the discrepancy derived from MLE ($\widehat{D}^{MLE} = \hat{S}_V - \hat{S}_H$)[26], contains noise and possibly bias. Therefore, as we discussed in the previous section, the basic

[25] In the following section it is assumed that two physical attributes are equal: $S_w{}^V = S_w{}^H$.
[26] Since this discrepancy is derived from direct sensory measurement, it is called sensory discrepancy.

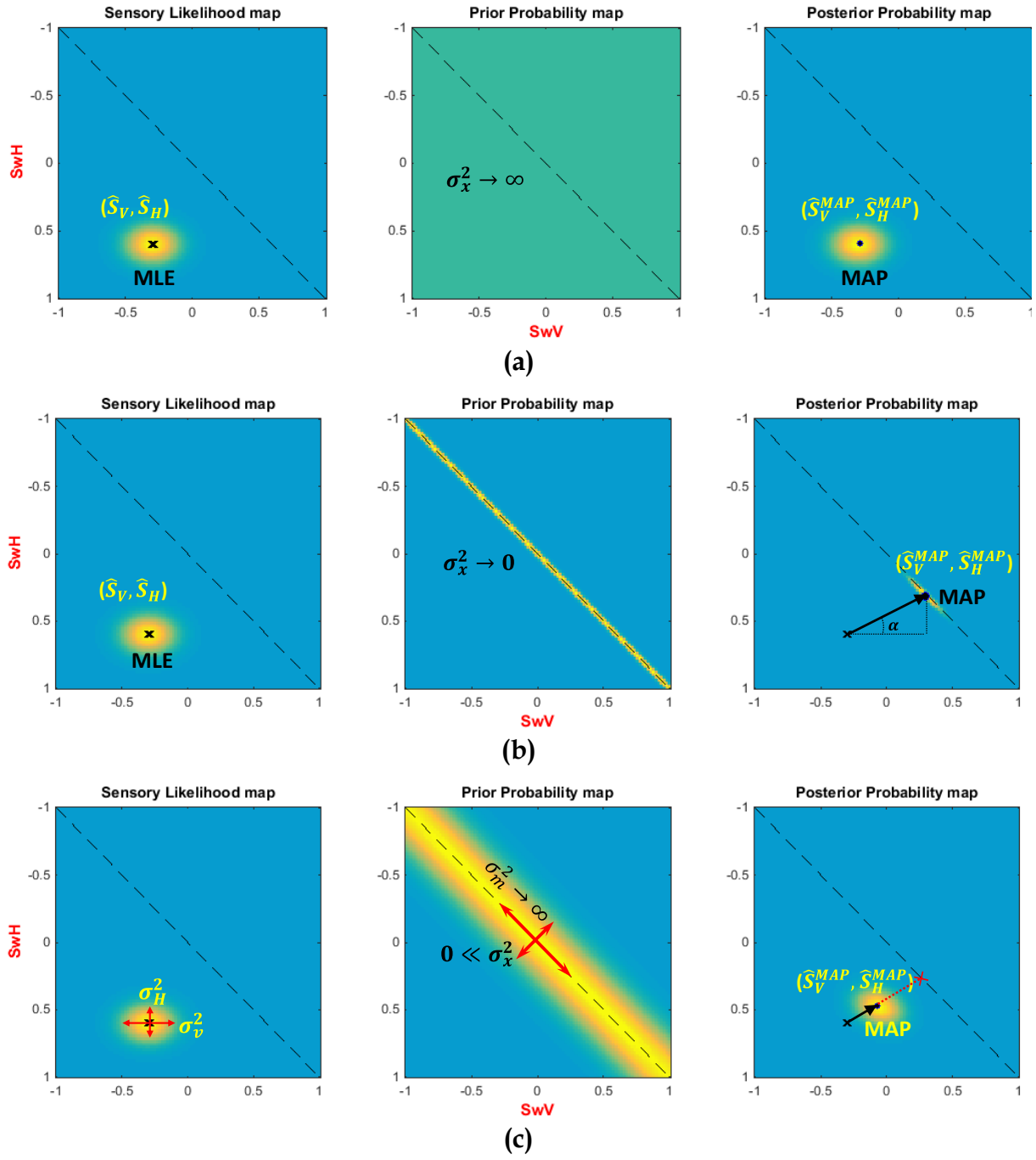


FIGURE 2-7

Illustration of Coupling-Prior model as a unified model of fusion and segregation. The joint distribution of sensory likelihood, prior and posterior are depicted for three different settings. (a) Full-Segregation scenario where $\sigma_x^2 \to \infty$. In this model MLE is identical to MAP and the prior reflects no relation between signals. (b) Full-Fusion model, in which $\sigma_x^2 \to 0$ and two signals are assumed to be perfectly correlated. (c) Partial-Fusion model as an intermediate model of full-fusion and full-segregation, where $0 < \sigma_x^2 < \infty$. MLE is indicated by black cross and the black arrow shows the drift from MLE to MAP estimate in each model of integration.

linear combination of equation (2-19) is error prone. This problem can be dismantled into two distinct problems:

- First, it is necessary to determine which portion of sensory discrepancy $\widehat{D}^{MLE}$ is caused by bias, or equivalently to estimate (B_V, B_H).
- Second, given the likelihood covariance Σ, and the estimated bias, it is necessary to integrate information in such a way that the final estimate exhibits minimum variance, along with reducing the bias.

Even though Bayesian Integration of the sensory likelihood and the prior probability, systematically minimizes the variance of the posterior distribution, the problem of bias estimation is a bit tricky. On the other hand, choosing a prior probability that properly models the statistical structure of the stimuli is not always easy [Conway and Christiansen 2006]. It is important to consider two important factors while we choose a prior. First, the sensory attributes that are subject to bias ($S_i = S_w{}^i + B_i$) represent a single physical event across different senses, thus they must be correlated[27]. In other words, there is a mapping between sensory attributes such that the occurrence of one can predict the other. In a sensory convergence scenario[28], it is often assumed that $S_w{}^V$ is identical to $S_w{}^H$.

However, since the prior belief is formed by sensory experiences rather than real-world signals, it cannot directly represent the statistics of the physical stimuli. The second factor in modeling the prior is the variance of joint-distribution that determines the variability of joint-occurrence, and it is influenced by inaccuracy [Ernst and Di Luca 2011]. This means the sensory conflict is most likely caused by the bias rather than noise. Larger the prior variance we choose in the model, greater the probability for discrepant signals to occur, and thus more likely a bias-driven conflict takes place. From another point of view, the variance of the prior reflects the belief in how precisely a single modal attribute can predict the other attribute. This notion highlights the prior variance as a parameter that quantifies the mapping uncertainty. To take into account the mentioned factors in a mathematical formulation, Ernst and Di Luca defined the following bivariate Gaussian prior, for a visual-haptic size discrimination task, where σ_x^2 tunes the mapping uncertainty:

$$P(S_V, S_H) = N\left(\hat{S}^P, \Pi\right), \Pi = R^T \begin{pmatrix} \sigma_m^2 & 0 \\ 0 & \sigma_x^2 \end{pmatrix} R,\ R = \begin{pmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{pmatrix} \tag{2-23}$$

Where, Π is the covariance matrix of prior joint distribution, and R is an orthogonal matrix that rotates the Cartesian coordinate by 45°. To make the prior joint probability

[27] This correlation is different from noise-correlation in each sensory modality.

[28] Sensory convergence is referred to the situation in which multiple sensory attributes are collected from a single physical phenomenon.

independent of the mean vector, distribution function is chosen to be diagonally symmetrical. So, σ_m^2 is chosen to be much greater than σ_x^2, e.g., at least ten times greater. Since this model of prior represents a coupling association between sensory attributes, it is called coupling prior. It is important to note that the coupling-prior does not necessarily reflect directly the structural statistics of the physical world, and that makes it slightly different from the prior distribution described in (2-16).

Given the prior and the likelihood joint distributions formulated in (2-22) and (2-23) and using Bayes rule, the posterior joint distribution can be obtained according to the following equation:

$$P(S_V, S_H|z_V, z_H) = P(z_V, z_H|S_V, S_H)\, P(S_V, S_H) \tag{2-24}$$

This joint distribution[29] gives rise to the final Maximum-A-Posterior estimate of the sensory signals $\hat{S}^{MAP}$. MAP can be acquired using equations (2-25) and (2-26), where $\hat{S}^P = (S_V^P, S_H^P)$ is the mean of joint prior distribution and $\hat{S}^{MLE}$ is the current sensory observation:

$$\hat{S}^{MAP} = \left(\hat{S}_V^{MAP}, \hat{S}_H^{MAP}\right) = W_\Sigma \hat{S}^{MLE} + W_\Pi \hat{S}^P \tag{2-25}$$

$$W_\Sigma = (\Sigma^{-1} + \Pi^{-1})^{-1} \times \Sigma^{-1}, W_\Pi = (\Sigma^{-1} + \Pi^{-1})^{-1} \times \Pi^{-1} \tag{2-26}$$

In Appendix A, a comprehensive mathematical analysis of the coupling-prior model, and how to derive (2-25) and (2-26), is reported. By expanding equation (2-26) and replacing it in equation (2-25), we have explained how $\hat{S}^{MAP}$ and its associated covariance matrix can be computed as a linear combination of current sensory measurement and the mean of coupling-prior. We have also shown that, for any arbitrary set of model parameters[30] the sum of W_Σ and W_Π is always equal to identity matrix (see Appendix A for more detail). Therefore, the linear integration model of equation (2-25) is similar to that of described in (2-19). The only difference is that, the scalar weights in (2-19) are replaced with the weighting matrices W_Σ and W_Π. At the following, we can see one instantiation of this model for the visual-haptic size discrimination task described in [Ernst and Di Luca 2011]. The derived intermediate equations can be found in Appendix A:

$$\hat{S}^{MAP} = \frac{1}{2\sigma_x^2 + \sigma_V^2 + \sigma_H^2} \left\{ \begin{bmatrix} (2\sigma_x^2 + \sigma_H^2) S_V + \sigma_V^2 S_H \\ \sigma_H^2 S_V + (2\sigma_x^2 + \sigma_V^2) S_H \end{bmatrix} + \begin{bmatrix} \sigma_V^2 (S_V^P - S_H^P) \\ \sigma_H^2 (S_H^P - S_V^P) \end{bmatrix} \right\} \tag{2-27}$$

Although, a-priori physical state vector $\hat{S}^P = (S_V^P, S_H^P)$ can be generally incorporated in this model, the final estimate will be independent of $\hat{S}^P$ in case the coupling-prior becomes diagonally symmetrical. This assumption implies an identical and in general a

[29] For the simplicity, the normalization factor of posterior distribution is not written.

[30] The introduced linear integration model of coupling-prior, includes two types of parameters: sensory prior variance σ_x^2, and sensory noise variance $\{\sigma_i^2\}$.

linear relation between sensory attributes. This assumption is often correct in most of the multisensory perceptual decision tasks [Ernst 2005] [Ernst 2007]. As a result, MAP estimate can be formulated according to the following equation:

$$\hat{S}^{MAP} = \frac{1}{2\sigma_x^2+\sigma_V^2+\sigma_H^2}\begin{bmatrix}(2\sigma_x^2+\sigma_H^2)S_V + \sigma_V^2 S_H \\ \sigma_H^2 S_V + (2\sigma_x^2+\sigma_V^2)S_H\end{bmatrix} \tag{2-28}$$

Now let us see the behavior of the model in two extreme cases where the mapping uncertainty (or equivalently prior variance) σ_x^2 approaches to infinity or zero:

$$\hat{S}^{MAP} = \begin{cases} \begin{bmatrix} S_V \\ S_H \end{bmatrix} & \text{if } \sigma_x^2 \to \infty \\ \begin{bmatrix} \frac{S_V\sigma_H^2 + S_H\sigma_V^2}{\sigma_V^2+\sigma_H^2} \\ \frac{S_V\sigma_H^2 + S_H\sigma_V^2}{\sigma_V^2+\sigma_H^2} \end{bmatrix} & \text{if } \sigma_x^2 \to 0 \end{cases} \tag{2-29}$$

Therefore, the behavior of the model is highly tied to the mapping uncertainty. But, the question is, what is the functional equivalence of this parameter? At the following paragraph, we have evaluated the characteristics of the coupling-prior model to explain how it can unify three processes of multisensory integration into a single framework. The key to this unification is the role of mapping uncertainty in the model outcome. The coupling-prior is an embodied model of mapping between real-world attributes which constrains the model of the integration and thereby instantiates different processes of integration. In FIGURE 2-7, we have illustrated three instantiations of the coupling-prior model with different values for σ_x^2. That leads to three basic functions: Full-segregation, Full-Fusion, and Partial-Fusion:

- **Full-Segregation:** In FIGURE 2-7-(a), the variance of the coupling-prior is set to infinity ($\sigma_x^2 \to \infty$) and forms a uniform joint distribution. This coupling represents a highly uncertain mapping between signals which are completely uncorrelated. Equivalently, this setup is associated with an observer that does not have any knowledge about the mapping function. Thus, one assumes there is no coupling between sensory attributes. Consequently, the posterior probability and thereby MAP estimate become identical to joint likelihood and MLE, respectively. As we can see in equation (2-29), if $\sigma_x^2 \to \infty$, MAP estimate ($\hat{S}^{MAP}$) approaches to MLE ($\hat{S}^{MAP} = (S_V, S_H)$). From another point of view, the flat coupling-prior implicitly implies a flat probability distribution for sensory discrepancy ($\hat{D}^{MLE} = \hat{S}_V - \hat{S}_H$), and thereby it is highly probable that a wide bias-driven conflict occurs. This setting leads to full-segregation process that introduces no benefit of variance-minimization into the final estimate.
- **Full-Fusion:** As another extreme case for σ_x^2, if we set it to zero, that gives rise to a perfect and certain mapping between sensory signals. This sharp mapping

constraints the posterior probability to include exclusively those pairs of sensory signals that lie along the mapping curve[31], i.e. $S_V = S_H$. This form of coupling-prior dictates full-fusion of signals which are assumed to be bias-free. By comparing equation (2-12) with (2-29), it is clear that Full-Fusion model is comparable with the one that we described in Section 2.3.1, known as *standard model* of cue integration. In FIGURE 2-7-(b), it is illustrated that the variance of MAP is maximally reduced but at the cost of a strong bias in MAP. The direction of the shift towards the sharp prior (black arrow in FIGURE 2-7) can be determined by the ratio of signal reliabilities, i.e. $\alpha = \tan^{-1}\left(\frac{\sigma_H^2}{\sigma_V^2}\right)$, for detailed mathematical analysis see Appendix A. In fact, the component of the shift vector[32] that corresponds to the less reliable modality, i.e. vision in FIGURE 2-7-(b), is greater than the one associated with more reliable signal, i.e. haptic.

- **Partial-Fusion** If we set a positive-definite value to the prior variance (i.e. $0 < \sigma_x^2 < \infty$, and much smaller than σ_m^2), the belief in presence of a coupling mapping between senses becomes positive non-zero. In this case the posterior distribution and thus MAP is located somewhere between that of calculated in two extreme cases. The direction of shift in the location of MAP estimate or equivalently the posterior distribution is identical to that of full-fusion case (the black arrow in FIGURE 2-7-(c) right). However, the length of the shift vector that explicitly exhibits the strength of fusion is shortened as compared to the Full-Fusion. Moreover, the variance of the posterior is also shrunk which is in fact the main advantageous outcome of the fusion. In FIGURE 2-8, we have derived and compared the principle components of posterior covariance matrices, in models with different values of σ_x. There are two important messages in FIGURE 2-8-(b): first, it is clear that the principle components of posterior covariance (σ_1^2, σ_2^2) are smaller than the corresponding components in likelihood distribution (σ_V^2, σ_H^2). This reflects the partial beneficiary feature inherited from fusion process. Secondly, as σ_x monastically shrinks, the posterior covariance components also monastically become smaller, and approach to zero. On the contrary, (σ_1^2, σ_2^2) becomes wider in such an extent to reach (σ_V^2, σ_H^2), as σ_x widens enough. The second fact shows the important role of σ_x in tuning the strength of fusion. This model of integration is an intermediate model of full-fusion and full-segregation and is called Partial-Fusion. Partial-fusion inherits the advantage of the fusion model (variance-minimization) while it avoids the undesired effect of the bias in the MAP estimate. This linear combination model provides an optimal balance between costs and benefits of integration, i.e. imprecision and inaccuracy minimization that are two competing factors in the fusion. Bresciani et.al confirmed how this model can predict the behavior of a human observer in a visual-haptic

[31] In multi-dimensional space, the relation function can be imagined as a manifold.
[32] Shift in MAP estimate with respect to MLE.

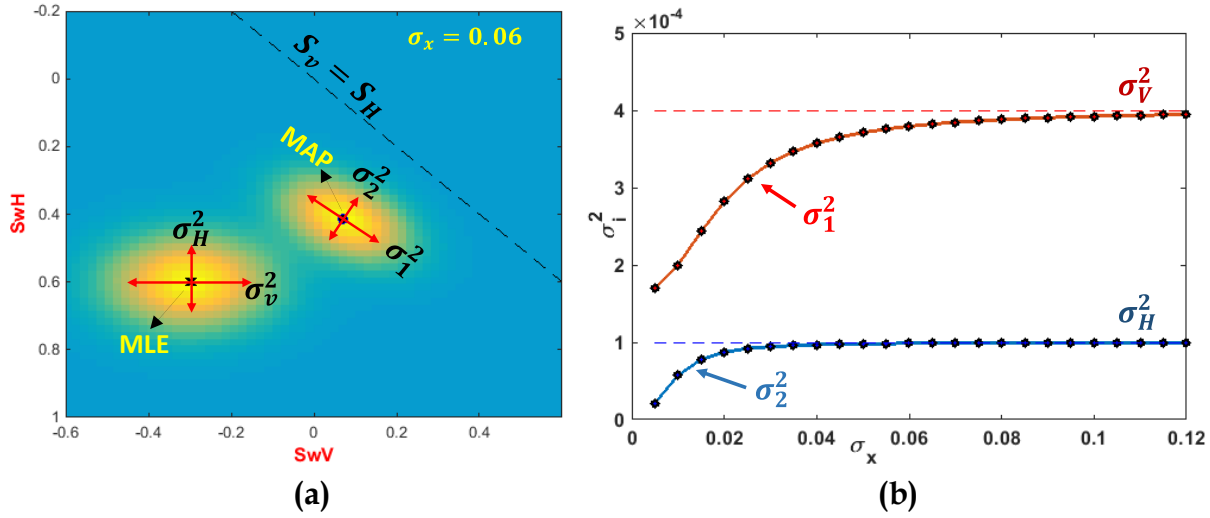


FIGURE 2-8

An illustration of partial variance-minimization in partial-fusion. **(a)** The Posterior and Likelihood joint distributions, and the principle components of their covariance matrices are depicted and compared, where σ_x is set to 0.06. The principle components of the posterior covariance matrix (σ_1^2, σ_2^2) are shrunk as compared to σ_V^2 and σ_H^2. **(b)** The graph of principle components of posterior covariance matrix, as a function of Coupling-Prior variance or equivalently mapping uncertainty σ_x. As is shown in this graph, the covariance components approach to sensory variances, as σ_x grows. On other hand, the posterior components are less than sensory variance in any circumstances. This is the main beneficiary of fusion process.

perceptual decision task [Bresciani et.al 2006]. Given $\sigma_x^2, \sigma_V^2, \sigma_H^2$ as system parameters and assuming $\sigma_x^2 \ll \sigma_m^2$, we have derived a linear system description of this model that combines sensory evidences $(\hat{S}_V, \hat{S}_H)$ to estimate $\hat{S}^{MAP}$. For more detail check Appendix A.

Thus far, we have introduced three processes of integration and a model that unifies them to potentially avoid the cost of fusion. However, the first problem of credit-assignment, i.e. bias estimation, is still not tackled. As we discussed, the portion of noise contribution in the current sensory discrepancy $\hat{D}^{MLE} = \hat{S}_V - \hat{S}_H$ can be reduced by the partial-fusion (see FIGURE 2-7 and FIGURE 2-8). Therefore, it is reasonable to think of the remaining discrepancy in MAP $\hat{D}^{MAP} = \hat{S}_V^{MAP} - \hat{S}_H^{MAP}$ as the best estimate of the actual discrepancy corresponding to the actual systematic bias. Having the MAP calculated so far, in the next section we will introduce a recursive algorithm to compute the best estimate of bias (B_V, B_H) and thereby to re-calibrate the perceptual system accordingly.

2.3.3.2 Sensory Recalibration and Integration Breakdown

2.3.3.2.1. Sensory Recalibration:

We discussed how to tackle the second problem of the credit-assignment in Section 2.3.3.1 to balance the benefit and cost of the integration. But, thus far, the introduced model does not cancel out the possible undesired effect of sensory bias in the final estimate. In the current section, we introduce a recursive algorithm that can be integrated within the coupling-prior model to solve the first problem of the credit assignment. This problem arises in case one of the sensory nodes exposes a persistent pattern of drift with respect to the physical property $S_w = (S_w^V, S_w^H)$. As opposed to the reliability, which can be directly reflected by sensory noise, the signals do not carry any direct information about the drift they carry. As a result, the bias should be estimated through trials and that requires a recursive process. The key idea is to use the available information observed at present and past trials to infer a rough estimate of an unknown bias. Then, at the next trial, this rough estimate will be updated as new information becomes evident. Obviously, integrating the new piece of evidence will increasingly enhance the quality of the final estimate at each trial. However, it is important to note that the assumptions and the constraints we define through the computational steps must describe the structure of the sensory world properly. Otherwise, the algorithm will not converge into a correct point. Similar to Kalman Filter (we will describe the mechanism of Kalman Filter in Section 2.3.4), the introduced algorithm can be interpreted as a two-layered recursive Bayesian inference which is elaborated at the following paragraph.

At the first layer, we will combine the sensory evidence observed at present time $S_t = (S_{V,t}, S_{H,t})$ with an existing coupling-prior $p(S_V, S_H)$ to compute the best current estimate of the physical state $\hat{S}_t^{MAP} = (\hat{S}_{V,t}^{MAP}, \hat{S}_{H,t}^{MAP})$. This layer of inference is introduced in the previous section as coupling-prior model of integration. In the left and the middle columns of FIGURE 2-9, we have demonstrated this process. Having the posterior estimate computed at time t and assuming $S_w^V = S_w^H$, the posterior discrepancy can be derived as:

$$\hat{D}_t^{MAP} = \hat{S}_t^{MAP} - \hat{S}_t^{MAP} = \left(S_w^v + \hat{B}_{V,t}^{MAP}\right) - \left(S_w^H + \hat{B}_{H,t}^{MAP}\right) = \hat{B}_{V,t}^{MAP} - \hat{B}_{H,t}^{MAP} \qquad (2\text{-}30)$$

As a result, all possible pairs of bias estimates $(\hat{B}_{V,t}^{MAP}, \hat{B}_{H,t}^{MAP})$ that can satisfy equation (2-30) will form a bias-likelihood distribution. This distribution is indicated by a blue line in the right column of FIGURE 2-9. To model the possible uncertainty in the estimated posterior discrepancy $\hat{D}_t^{MAP}$, the bias-likelihood contains Gaussian noise and that is, the blue line becomes blurry in FIGURE 2-9.

At the second layer, we combine the derived bias-likelihood with a pre-defined bias-prior to compute the bias-posterior (see right column of FIGURE 2-9). Ghahramani et.al proposed that the contribution of each modality in a sensory conflict should be proportional to its variance [Ghahramani et.al 1997]. This notion suggests more credit in

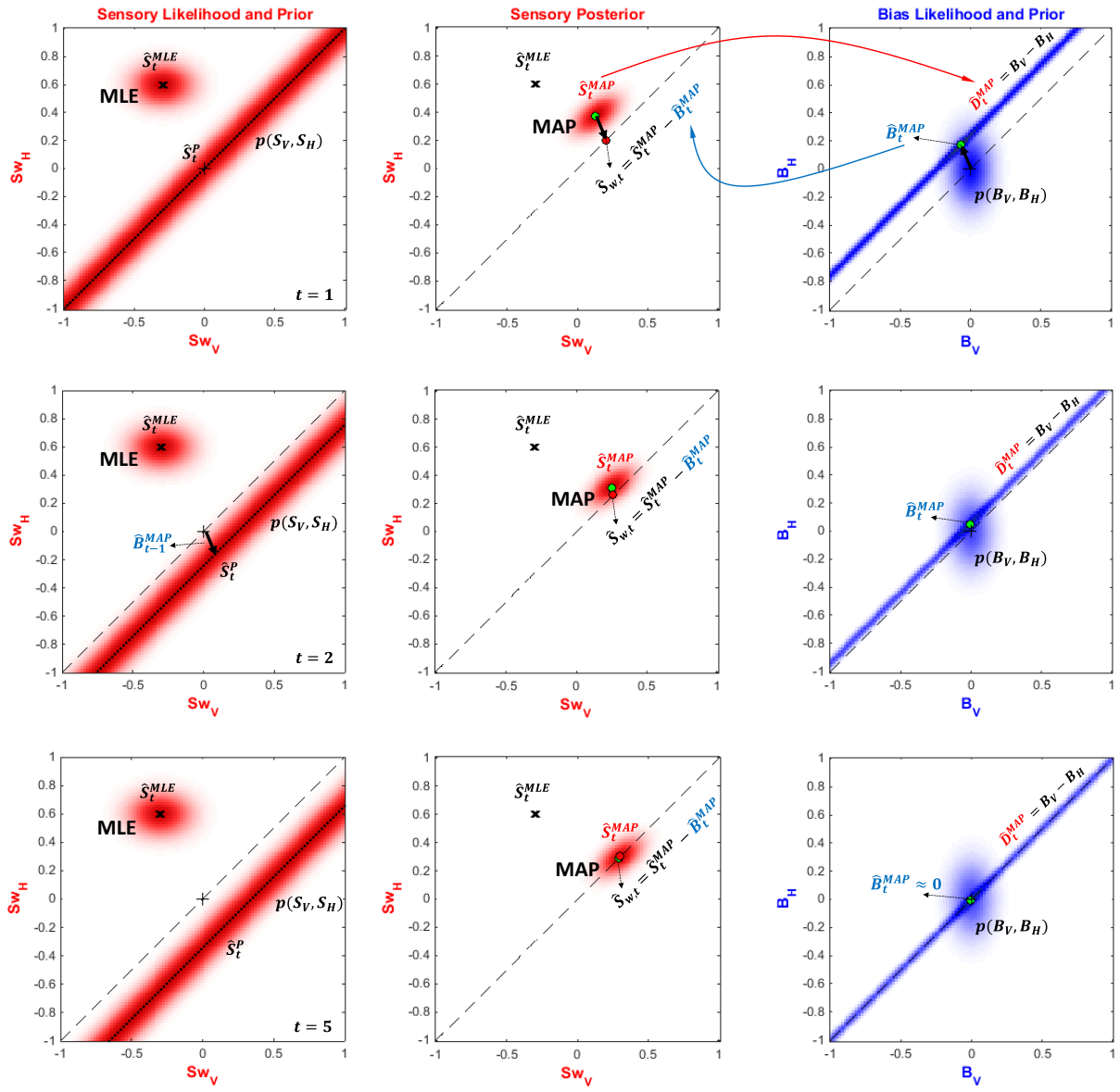


FIGURE 2-9

Iterative process of Sensory Recalibration and Remapping in Coupling-Prior model in a visual-haptic size estimation task. In the left column, red blob represents the hypothetical sensory joint-likelihood which is assumed to be constant within iterations, and the blurry line represents the coupling-prior. In the middle column, the estimated sensory posterior (red Gaussian blob), MAP estimate $\hat{S}_t^{MAP}$ and estimated physical state variable $\hat{S}_{w,t}$ are shown. The discrepancy derived from MAP estimate $\hat{D}_t^{MAP}$ is used to create the bias-likelihood. In the right column the bias-likelihood is indicated by a blurry blue line and represents all possible pairs of bias values that can potentially cause the estimated posterior discrepancy. By combining the bias-likelihood with a pre-defined bias-prior (blue Gaussian blob in right figures) to estimate bias posterior. Then the estimated bias at current step is used to update the coupling-prior or sensory mapping at the nest iteration (middle row).

bias estimation for the sensory signal with higher variance. This strategy is sub-optimal and cannot be valid, because the variance gives no information about the probability of the exposing bias. A simple and reasonable way to choose a suitable bias-prior is to consider the probability of how often each single node is biased in past experiences. For example, if the haptic signal caused more conflicts in the past, then, it is more probable to cause conflict in the current situation. In a visual-haptic size discrimination task, Ernst and colleagues suggested a Gaussian bias-prior in which the haptic principle component of its covariance matrix is wider (the blue Gaussian blob in FIGURE 2-9) [Burg et.al 2008] [Ernst and Di Luca 2011]. This reflects the fact that the haptic signal more frequently shows bias compared with visual signal. Having the bias-likelihood and bias-prior at hand, we can solve the second problem of credit-assignment by computing bias-posterior. Thereafter, in the next iteration we can update the coupling-prior that represents the mapping between signals. This remapping process is done by shifting the coupling-prior according to the estimated bias vector:

$$p(S_{V,t+1}, S_{H,t+1}) = p(S_{V,t+1} - \hat{B}_{V,t}^{MAP}, S_{H,t+1} - \hat{B}_{H,t}^{MAP}) \quad (2\text{-}31)$$

For instance, in left column of FIGURE 2-9 (b), the coupling-prior is shifted by the bias estimated in the previous iteration. This also results in an enhanced quality of estimation at each iteration, both for the bias and the physical attribute $\hat{S}_{w,t}$. Now we can exclude the estimated bias component from MAP estimate:

$$\hat{S}_{w,t} = \hat{S}_{t}^{MAP} - \hat{B}_{V,t}^{MAP} \quad (2\text{-}32)$$

In FIGURE 2-9, we have described the mechanism of iterative recalibration process through a simple example. More interestingly, at each iteration of the algorithm, the estimated bias converges to zero and thereby the MAP estimate becomes identical to the physical estimate. This is the consequence of iterative remapping process that corrects the prior belief in relation between sensory modalities. As is shown in FIGURE 2-9, the initial coupling-prior is assumed unbiased, but as the algorithm develops, the prior reflects the estimated bias.

2.3.3.2.2. *Integration breakdown:*

Sometimes the sensory conflict is not necessarily due to the bias and it is possible that the sensory attributes are caused by separate sources. In this case, the observer should first infer present situation, and thereby fuse the signals into a single estimate or segregate them as irrelevant descriptive features. Even in some cases that the signals belong to a single physical source, they might exhibit a large conflict in time or space. Human observer is able to break down the fusion in these multisensory scenarios [Wallace et.al 2004] [Roach et.al 2006] [Körding et.al 2007] [Shams 2012]. But, the question is how a model can take it into account? As is stated in Section 2.3.3.1, the parameter that determines the underlying process of integration in coupling prior model is a-priori

variance σ_x^2. Thus, the present sensory discrepancy does not influence the underlying process of integration. For large sensory conflict, whether it occurs in time, space, or sensory coordinate, the integration process must break down. Otherwise, the estimation is not accurate, and the perceptual system is not robust. The way coupling-prior model incorporates integration breakdown is synonymous to that of segregation using an embodied flat prior. The desired outcome is to exclude the discordant or discrepant[33] sensory signals from fusion that exceeds a specific temporal or spatial threshold. To incorporate this functionality, Roach et. al introduced a Gaussian-like prior with a heavy tail. The shape of this prior is similar to a Gaussian, but it does not approach to zero for the values far from the center. Instead, it keeps a uniform non-zero probability for those sensory pairs that exceeds the threshold [Roach et.al 2006]. In other words, this model of prior is a piecewise linear combination of a uniform and Gaussian joint distribution. As a result, the perceptual system can still perform partial-fusion as long as the detected sensory conflict falls into the Gaussian-side, otherwise the integration breaks down because of the increased influence of the flat-side.

In chapter 5, we will introduce a sophisticated and hierarchical Bayesian inference model called Causal Inference that accounts for integration breakdown.

2.3.4 Dynamic Bayesian Models, Kalman Filter & Particle Filter

As we discussed in chapter 1, twisted interplay of perception and motor control is essential for survival [Wolpert & Ghahramani 2000]. On the other hand, one of the key processes in multisensory integration is the intervention of motor system in perceptual system through which the understanding of the sensory world might change. When we take an action to manipulate and to interact with our environment, we change the internal state of our body as well as the external state of the world. Moreover, regardless of taking any actions, some of the physical stimuli themselves are not static and their attributes are changing in time, e.g. moving objects, changing light intensity, changing in posture or pitch, etc. Therefore, the perceptual system must encounter these forms of dynamics.

In Section 2.3.1 and 2.3.2, we introduced the standard model of the Bayesian fusion to optimally integrate multiple senses into a single estimate. We have also introduced a unified linear model of the Bayesian integration that can integrate multiple perceptual functions within a single framework (see Section 2.3.3). These models can optimally operate on a static multisensory set-up. The key to Bayesian inference is to combine the prior belief in a set of hypotheses (e.g., internal or external state variables) with current evidence to compute the probability of the occurring hypothesis. But, how can it be formulated for a dynamic perceptual task? To drop this notion into a theoretical

[33] Sensory discordance is usually defined as the sensory conflict in time and space. On the other hand, sensory discrepancy is the conflict within sensory coordinate.

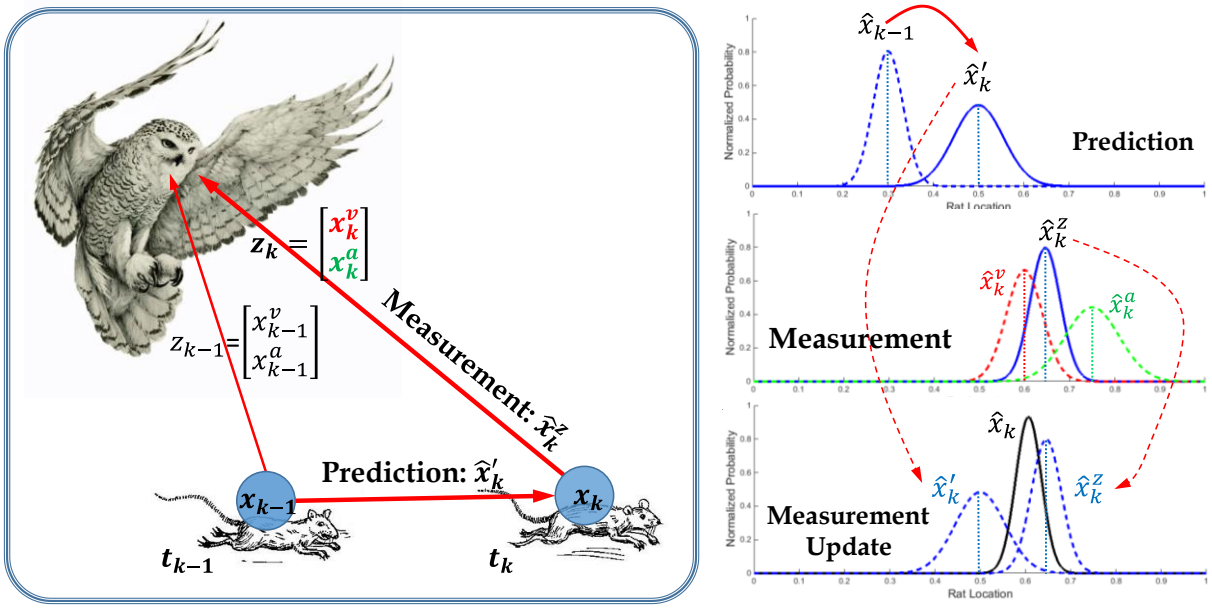


FIGURE 2-10

Left: General mechanism of a simple dynamic Bayesian inference model in a real-world multisensory integration problem. A snowy owl is going to hunt a moving rat, while the perceptual system accounts for the best estimate of the target's location at each time steps. The location of the rat, that can be seen as external state variable (x_{k-1} and x_k), is hidden and not directly observable. At each time step, there is only a set of multisensory measurements of the state variable available (z_{k-1} and z_k). So, the owl's perceptual system should use optimally the available information to compute $\hat{x}_k$ to guide the motor system to the target. The red arrows represent the dependency of signals and the way information flows in time.

Right: Three graphs represent the basic computational phases that should be undertaken within the dynamic Bayesian integration. In the prediction phase (top graph), given the previous estimate of the state variable $\hat{x}_{k-1}$, the animal can predict the prey's location before picking up the sensory attributes at present time. This prediction is indicated by a Gaussian profile centered at x_k'. In the measurement phase (middle graph), the sensory attributes x_k^v and x_k^a are captured by the predator's nervous system and are combined to give an estimate of the current state variable $\hat{x}_k^z$. This is synonymous to that of described in Section 2.3.2. Eventually, in the measurement-update phase (lower graph), the predicted state variable x_k' is combined with the sensory measurement-driven estimation of the state variable $\hat{x}_k^z$ to compute an estimate of the rat's location at present time $\hat{x}_k$. This variable will be passed to the prediction phase of the next time step t_{k+1}. This process continues while the animal catches the rat. In this example priori and posterior beliefs are assumed to be Gaussian. But in general, this dynamic Bayesian scheme can be generalized for any arbitrary density functions.

framework, let us analyze a real-world example in which a snowy owl wants to hunt a rat (see FIGURE 2-10). To take any actions toward the moving target, the owl's perceptual system should give an estimate of rat's location[34] to its motor system at each time step. This variable which is called external state variable (x_k in FIGURE 2-10) is not directly observable and the nervous system has only access to a visual and acoustic measurements of that at each time step: $z_k = (x_k^v, x_k^a)$. Subscript k denotes a time instant t_k in the dynamic problem. In the previous sections, we summarized the models that can combine the visual and acoustic attributes to give an optimal estimate of a hidden variable, i.e. $\hat{x}_k^z$ in FIGURE 2-10. Since the target is moving and given an estimate of the previous location of the target $\hat{x}_{k-1}$ (blue dashed Gaussian profile in top graph of FIGURE 2-10), the nervous system is able to internally predict the current location before picking up the sensory evidence at t_k. This prediction value is indicated by $\hat{x}_k'$ in FIGURE 2-10 and can be considered as the prior belief in a possible location of the rat. Now at t_k when both the predator and the prey have changed their positions (that corresponds to internal and external state variables respectively), the provided visual-acoustic sensory signals give the owl a new evidence about the current location of the rat. This measurement-driven estimate of the state variable is indicated by $\hat{x}_k^z$ in FIGURE 2-10 and is computed using a standard Bayesian fusion algorithm (see the middle graph of FIGURE 2-10). Finally, the animal combines the prior prediction, and the sensory-driven estimate together to update its internal belief in rat's location. A sequence of this prediction-measurement computation gives rise to an accurate and reliable way of tracking a moving target, and thereby generating a proper sequence of action in order to guide the animal towards the prey.

Now let us put the spotlight on a mathematical description of one the most well-known dynamic Bayesian inference model, Kalman Filter, in the context of multisensory integration.

2.3.4.1 Kalman Filter:

Kalman Filter (KF) is known as one of the mostly used dynamic Bayesian inference models [Grover and Hwang 2012]. It is interesting to note that, back in 1969 the Apolo-11 mission used KF to estimate the trajectory of the spacecraft towards moon [Grewel and Andrews 2010]. This algorithm is in fact a linear stochastic differential equation[35] with first-order dynamics that characterizes the state of a system through a defined set of variables, called state variables. State variables hold the status of the system in time - whether continuous or discrete[36] – and forms a state-space in which a single state variable

[34] The location of the target can be defined within body-centered coordinate system.

[35] A differential equation in which the variables are random. In discrete time-space, it rather forms a stochastic deference equation.

[36] In this section the discrete form of Kalman filter is formulated. However, the continuous form of this algorithm is conceptually similar to discrete Kalman filter.

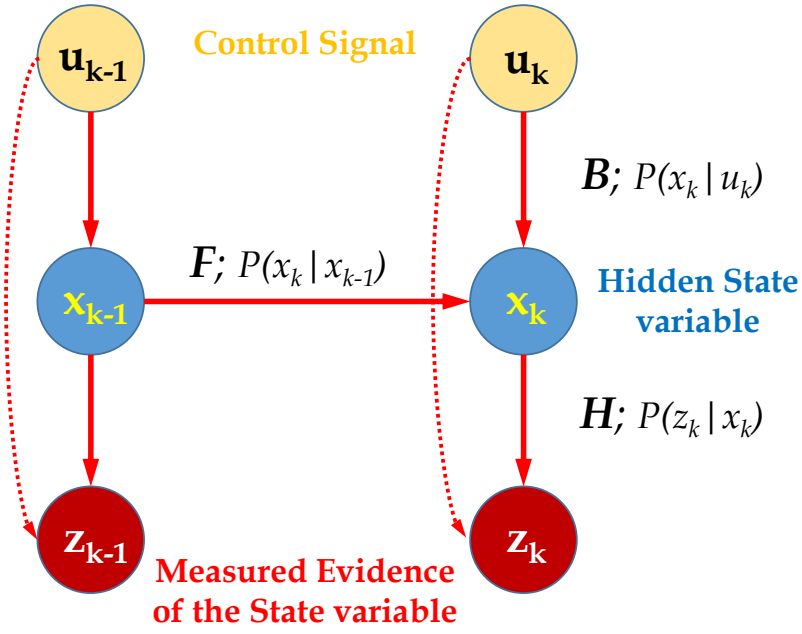


FIGURE 2-11

General scheme (Bayesian graph) of a Discrete Kalman Filter Algorithm. The orange nodes represent the input or control signals u. Blue nodes show the state variables x_k that are not directly observable and should be estimated from sensory evidences. And the red nodes, z represent the sensory measurements or sensory evidence of the hidden state. Arrows depict the conditional relationship and dependency between random variables. It is assumed that x_k is just conditionally dependent on previous state variable x_{k-1}; and z_k is also independent of sensory measurement in previous states $z_{i=0:k-1}$. A dynamic system under these assumptions is called a Markovian process (for more detail see [Ghahramani 2001]). KF is also a linear quadratic Markovian process. F is the transition mapping that relates the current state x_k to previous state x_{k-1}. B relates the input signals to the state variables. And H is a measurement mapping that models the relation between state variables and sensory measurements at each time-step. Mapping functions can be also seen as conditional probability functions. Indices k and $k-1$ denote the current and previous time steps respectively. The dashed line represents the possible mapping between input signal and sensory measurement, but we have excluded such a mapping in this section.

corresponds to an intermediate or a single output variable of the system. This way of modeling of a dynamic system resembles the state-space notion in control theory which simplifies analysis and control of that system [Grover and Hwang 2012]. Similar to a basic dynamic Bayesian model described at the first paragraph of this section, KF is also composed of two phases: prediction phase, and measurement-update phase.

The prediction phase is a linear mapping from the previous state to the current state[37] as is formulated in the following equation. Assuming a system with n state variables, F is a $n \times n$ matrice which relates x_k to x_{k-1}, and u is input control signal (see FIGURE 2-11):

$$x_k = Fx_{k-1} + Bu_k + w_k \tag{2-33}$$

Sometimes the observer (the perceptual system) needs to take an action and to bring the state of the system (whether internal or external) into a desired state. That requires a linear mapping between input control signals u and state vector x. u is often assumed to be deterministic. This is modeled by matrix B in (2-33). w_k is a vector of n random variables whose elements represent the governing Gaussian noise process in each state variable. As the algorithm evolves, along with a-priori estimate of the state vector $\hat{x}'_k$, the covariance matrix of the process noise can also be predicted in prediction phase i.e. P'_k:

$$\hat{x}'_k = F\hat{x}_{k-1} + Bu_k \tag{2-34}$$

$$P'_k = F\hat{P}_{k-1}F^T + Q \tag{2-35}$$

Where, Q is the initial covariance matrix, P'_k is the predicted covariance matrix at the current time step, and $\hat{P}_{k-1}$ is the estimated covariance matrix at the previous time step. The predicted state vector $\hat{x}'_k$ and the covariance P'_k are in fact a-priori estimates of the state vector and its respective covariance. These predictions will be updated in the measurement phase.

In the measurement phase, the sensory evidence described by $z_k \in R^m$ is related to x_k by using $m \times n$ matrix H (see FIGURE 2-11). The sensory noise is modeled by an additive Gaussian noise v_k with covariance matrix R:

$$z_k = Hx_k + v_k \tag{2-36}$$

Having the sensory evidence measured and given (2-34), (2-35) and (2-36), we can drive the posterior probability of the hidden state that causes the current sensory evidence. It is important to note that KF is a Markovian estimator. That means z_k is assumed to be statistically independent of $\{(x_i, z_i)|i = 0:k-1\}$ and x_k should be independent of $\{x_i|i = 0:k-2\}$. In FIGURE 2-11, the Bayesian graph of KF is illustrated. Consequently, from the Bayesian sense, the prior is given by (2-33) and the likelihood is given by the sensory measurement:

$$P(x_k|z_k) = P(z_k|x_k)P(x_k|x_{k-1}) \tag{2-37}$$

[37] Sometime the next state is desired to be predicted, given the current state of the system. Conceptually they are the identical.

As a result of linear Gaussian process, the maximum posterior probability of equation (2-37) will give rise to the following linear estimation of the state vector in current time step $\hat{x}_k$. For more details see [Grover and Hwang 2012]:

$$\hat{x}_k = \hat{x}'_k + K_k(z_k - H\hat{x}'_k) = \hat{x}'_k(I - K_k H) + K_k z_k \quad (2\text{-}38)$$

$$K_k = P'_k H^T (H P'_k H^T + R)^{-1} \quad (2\text{-}39)$$

$$\hat{P}_k = (I - K_k H) P'_k \quad (2\text{-}40)$$

Where, K_k is a $n \times m$ matrix that is called *Kalman gain*, and $\hat{P}_k$ is the estimated state covariance. Equations (2-34) and (2-35) rule the prediction or state-transition phase of the algorithm, while equations (2-38), (2-39), and (2-40) are the governing equations in measurement-update phase. In practice, *F*, *B*, and *H* might change in time, but here we have assumed them constant.

In the right-hand side of equation (2-38), term $(z_k - H\hat{x}'_k)$ is called *sensory residual* and it reflects the error between actual sensory measurement and the expected sensory measurement based on predicted state vector $\hat{x}'_k$. The compensation term of the predicted state is the *sensory residual* weighted or equivalently transformed by *Kalman gain*. In other words, the contribution factor of the sensory evidence in the final estimate of the state vector is determined by K_k. More interestingly, if the sensory covariance *R* approaches to zero in (2-39), Kalman gain becomes equal to H^{-1} and $\hat{x}_k = H^{-1} z_k$. That means, a perfectly precise sensory evidence results in zero contribution of a-priori estimate $\hat{x}'_k$ in equation (2-38). On the other hand, as the a-priori state covariance, i.e. P'_k[38], approaches to zero, K_k also approaches to zero, and that leads in zero contribution of sensory evidence.

KF is an iterative process in which the output of the previous iteration is the input to the next (FIGURE B-1). This style of information fusion allows the filter to converge towards a more accurate estimate and to cancel out the perturbations caused by intrinsic noise or systematic bias.

2.3.4.2 *Extended Kalman Filter:*

The state transition and measurement mappings (e.g. *F* and *H* in FIGURE 2-11) might be non-linear. In this case, the KF algorithm is referred as Extended Kalman Filter in which the nonlinear mappings are linearized using Tylor expansion around a-priori state variable and sensory observation. As a result, the respective elements of the Jacobian matrix of nonlinear mappings, will be replaced with *F*, *B* and *H* matrices. For example, for a dynamic system with an arbitrary transition function $x_k = f(x_{k-1}u_{k-1}, w_{k-1})$ and

[38] That means a perfectly precise and error-free state prediction.

measurement function $z_k = h(x_k, v_k)$, the equation for prediction phase will be as follows:

$$\hat{x}'_k = f(\hat{x}_{k-1}, u_k, 0) \tag{2-41}$$

$$P'_k = F_k \hat{P}_{k-1} F_k^T + W_k Q W_k^T \tag{2-42}$$

Where, $F_k[i,j] = \frac{\partial f_i}{\partial x_j}(\hat{x}_{k-1} u_k, 0)$ and $W_k[i,j] = \frac{\partial f_i}{\partial w_j}(\hat{x}_{k-1}, u_k, 0)$. The measurement-update phase is also governed by the following equations, Where $H_k[i,j] = \frac{\partial h_i}{\partial x_j}(\hat{x}'_k, 0)$, and $V_k[i,j] = \frac{\partial h_i}{\partial v_j}(\hat{x}'_k, 0)$:

$$\hat{x}_k = \hat{x}'_k + K_k\big(z_k - h(\hat{x}'_k, 0)\big) \tag{2-43}$$

$$K_k = P'_k H_k^T (H_k P'_k H_k^T + V_k R V_k^T)^{-1} \tag{2-44}$$

$$\hat{P}_k = (I - K_k H_k) P'_k \tag{2-45}$$

The fundamental pitfall of linearization is the fact that the transformed random variables are no longer Gaussian-like and thus the EKF is suboptimal as compared to the linear KF. However, EKF is still known as a simple and reasonably suboptimal approximation of the Bayes rule [Grewel and Andrews 2010]. In Appendix B, I have analyzed a realistic case study in detail to describe how to design the parameters of an EKF. The case study is a tracking problem in which the measurement function is a nonlinear mapping. In Appendix B it is shown that EKF can still effectively model the behavior of a nonlinear dynamic system, even though the transformed sensory signals are not normal.

2.3.4.3 Particle Filter:

The optimality of Kalman filter is guaranteed based on two assumptions: first, the noise process of the sensory measurement and the state noise process must be both additive Gaussian-like. Second, the dynamics of the system should be linear. For instance, in FIGURE 2-10, the location of the rat in the retinal and acoustic coordinates of the owl sensory system has an elliptical shape. As a result, it can be fit into a 2D Gaussian profile and the owl can model the location of the prey as a random variable with gaussian-like distribution. On the other hand, the state transition and the measurement mapping can be modeled by a linear matrix transformation. There are cases in which the environment does not fit into a linear model or the noise is not governed by a Gaussian-like process. For example, in a prediction problem where the possible state variable can just fall into one of two crescent-shaped regions, or the sensory evidence may have a very long tail as

TABLE 2.1

Flow of standard particle filter algorithm

The objective:

Approximate the posterior probability function: $f(x_k|\{z_j\}_{j=0}^{k},\{x_i\}_{i=0}^{k-1})$

Assumptions:

The dynamic process is constrained by *Markovian* assumptions.

State transition density function: $x_k \sim f(x_k|x_{k-1},u_{k-1})$.

Measurement density function: $z_k \sim h(z_k|x_{k-1},u_{k-1})$.

Step 0 (initialization**):**

Set the number of particles: *N*.

Initialize particles by drawing them randomly $x_0^i \sim P(x)$ for $i = 1,2,3,\dots,N$.

Initialize particle weights uniformly: $w_0^i \sim \frac{1}{N}$ for $i = 1,2,3,\dots,N$.

Do

For *t = 1* to *k*

Step 1: Given z_t, Draw and Normalize $w_t^i = h(z_t|x_{t-1}^i,u_t)$ for $i = 1,2,3,\dots,N$.

Step 2: Resample $\{x_t^i\}_{i=1}^N$ from $\{x_t^i,w_t^i\}_{i=1}^N$

Step 3: Propagate x_t^i by drawing x_{t+1}^i from $f(.|x_t^i,u_{k+1})$ for $i = 1,2,3,\dots,N$.

End-For

End Do

End of the Algorithm

opposed to the Gaussian profile. The performance of KF will be drastically dropped in these cases. Even EKF approximation cannot restore the required optimality especially when the posterior probability function becomes non-Gaussian.

The key arising issue that should be encountered is in fact modeling an arbitrary non-Gaussian posterior function, given a priori state transition function *F*, and sensory likelihood; see equation (2-37). Instead of modeling a stochastic process with fitting a standard density function into it and computing its respective covariance and mean, the *law of large numbers*[39] in probability theory enables an alternative way to represent any arbitrary density functions: *Monte-Carlo numerical approximation*. This is the main difference and the prominent benefit of the particle filter as compared to Kalman filter. However, it comes with the cost of higher computational effort. Let us assume that we have an approximation of $P(x_k|x_{k-1},u_k)$, so that we can draw random samples $\{x_k^i|i = 1{:}\,N\}$ or to be literal, particles from the process. Now we can formulate a quasi-prior distribution as a sequence of Dirac functions centered at particles. This is the prediction

[39] The law of large numbers in statistics states that: as the number of identically distributed, randomly generated variables from a stochastic process increases, the frequency of samples, the average, and the numerical variance, asymptotically approach to the respective theoretical parameters, i.e. probability density, mean, variance.

phase of PF algorithm. In the update-phase, having z_k measured, one can compute the residual $e_k = \{z_k - h(x_k^i, u_k)\}_{i=1}^N$. And then, the associated likelihood weights for each possible state x_k^i can be drawn and normalized respectively. We represent these weights by $\{w_k^i | i = 1{:}N\}$. Finally, the priori state prediction is combined with likelihood weights according to Bays rule:

$$P(x_k|z_k) = \sum_{i=1}^N w_k^i \, \delta(x_k - x_k^i), \ \sum_{i=1}^N w_k^i = 1 \tag{2-46}$$

in which $\delta(x_k - x_k^i)$ is Dirac delta function. In the sequel, the posterior particles should be propagated to the next step, by drawing $x_{k+1}$from $f(.|x_k^i, u_{k+1})$. As the number of samples grows and algorithm evolves in time, this estimate becomes a better approximation of the posterior probability function, and thereby that provides a more optimal solution. A single defined particle is a possible state that the system might fall in, and its respective weight represents how likely the pair of $\{x_k^i, z_k\}$ can take place. In TABLE 2.1, the flow of a standard form of PF algorithm is summarized.

Even though Kalman filter acquires much lower computational requirements, it is less flexible in terms of modeling the dynamics of a system. Once we collect sufficiently large number of samples, a particle filter enables us to handle almost any type of models. However, as the size of state vector increases, it is possible that one particle dominantly takes over the prediction. As a result, some areas of the state space would not contribute in the process of inference while we allocate computational resources to them. This phenomenon is called *degeneracy* and that can be solved by enlarging the number of particles. Another way to overcome this problem is to resample the particles from an effectively chosen prior called *Importance density*. The mitigation of *degeneracy* phenomenon is still an active area of research [Gustafsson 2010].

2.3.5 Integration of Utility Function within Action-Perception loop

To program and to send a proper motor command, CNS needs to choose a motor output from a set of possible motor commands. Therefore, firstly the brain needs to achieve a coherent perceptual understanding of the sensory world (external state) and internal state of the body, then, creates a mapping from the estimated states to the proper actions. Each action that is taken can change the internal state of the body and possibly the external state of the world. Thus, the perceptual system should compensate the sensory consequences of the action, and changes in the real-world state. In Section 2.3.4, a dynamic model of perception is introduced that can account for such a cognitive need. Like the introduced models of sensory perception, action-generation which is anchored in decision making can also be systematically modeled within a probabilistic Bayesian framework [Körding & Wolpert 2006]. But, the process of perception is slightly different

in terms of computation, in which the mapping from the sensory-state vectors to the motor-state variables is usually mediated by high-level factors, e.g. intention or goal of action. Our daily tasks to interact with our world including localization, navigation, and the voluntarily movements[40] are associated with a goal that should be accomplished within a sequence of actions. These actions are usually programed[41] in Pre-frontal Cortex that receives information from uni-sensory and poly sensory areas of the cortex, e.g. early visual cortices, V4/V5, A1/A2, and Parietal Cortex. These areas of the sensory cortex are known to create a hierarchical perceptual mapping of the real-world stimuli within single-modality coordinates (e.g., eye-centered), and cross-modality coordinates (body-centered, head-centered). On the other, hand pre-frontal cortex, a model of cortical computing that conducts the whole process, should include the task objective (goal), besides the process of multisensory combination and estimating the hidden state of the world. The task objective shapes the *Gain/Loss function* which is associated with the action. This function which is called usually *utility function* quantifies the desirability of the action's outcome. Given a set of possible actions to take, CNS should internally determine the consequence of the action and the associated benefit or loss of that action.

To incorporate the *utility function* and *goal* in a computational framework, Körding and Wolpert proposed a Bayesian framework that can describe the main characteristics of action generation in human subjects [Körding & Wolpert 2006]. The key feature of Bayesian computation is optimality. This framework enables a systematic way of combining the sensory-motor evidence (belief) with our goals in order to make an optimal or equivalently rational decision. Rationality is defined in conjunction with the utility function or the cost function. Given an action to take, its outcome can be combined with the associated utility values. The most favorable action is the one that maximizes the expected utility, U:

$$E(U) = \sum P\left(Outcome^i \middle| action\right) U(Outcome^i) \tag{2-47}$$

To quantify how desirable (good/bad) the action is, we need to associate each single action – of a set of finite actions – with utility function (or equivalently cost function). In general, desirable movement is the one that consumes less energy. Some ethological cost functions include *movement smoothness* and *accuracy* [Cruse et.al 1990] **[**Balasubramanian et.al 2015**]**. These utility or cost functions can describe target-directed actions. In reinforcement learning the reward is interpreted as the utility function.

[40] In general there are four types of movements in human, and mammals: *Reflexes* which is automatic triggered in response to salient sensory stimuli, e.g. eye-blink; *Postural* movements that is used to maintain an upright position with respect to gravity; *Rhythmic* movements, e.g. walking, chewing; and *Voluntary* movements that is entirely initiated within CNS and are associated with a goal.

[41] Motor programming sometimes is referred as motor planning in the literature.

Chapter 3

Cooperative Event-Fusion for Depth Estimation by using Stereoscopic Silicon Retinas

"It's not enough to be busy, so are the ants. The question is, what are we busy about?" - **Henry David Thoreau (1817-1862 AD)**

3.1 The problem of Stereoscopic Image Fusion for Depth Estimation

Depth perception is a crucial skill of animals and humans for survival. A predator is able to catch the prey in a very fast time scale, cats can jump onto the table, and birds can land on narrow edges and catch insects at the first attempt. These abilities and in general all behaviors that involve moving around in an environment require a precise estimate of how far away a visual object might be. In general, there are two major types of cues in the environment that help animals in depth perception: external cues that are captured just by using one eye (monocular cues), e.g. perspective or relative size of the objects in the scene; and internal cues that rely on the physiological processing of the visual stimuli. Neurons in the visual cortex can compute distance using motion parallax cues and relative movement of retinal images [Bruce et.al 2003]. The most important internal cue is retinal disparity, which is defined as the difference between positions of the objects on the retinal images. This cue is the anatomical consequence of the eyes' positions on the face. The ability to use retinal disparity in depth perception is known as stereopsis in vision and still is an active research field (see FIGURE 3-1).

Following the fact that many basic aspects of the human visual processing system have been discovered in recent years, VLSI technology addresses emulating brain circuitry by introducing new microchips and brain-like information processing circuits, e.g. Dynamic Vision Sensors, Silicon cochlear, and massively distributed processors (SpiNNaker) [Indiveri and Douglas 2000] [Wen and Boahen 2009] [Ferber and Brown 2009]. One open

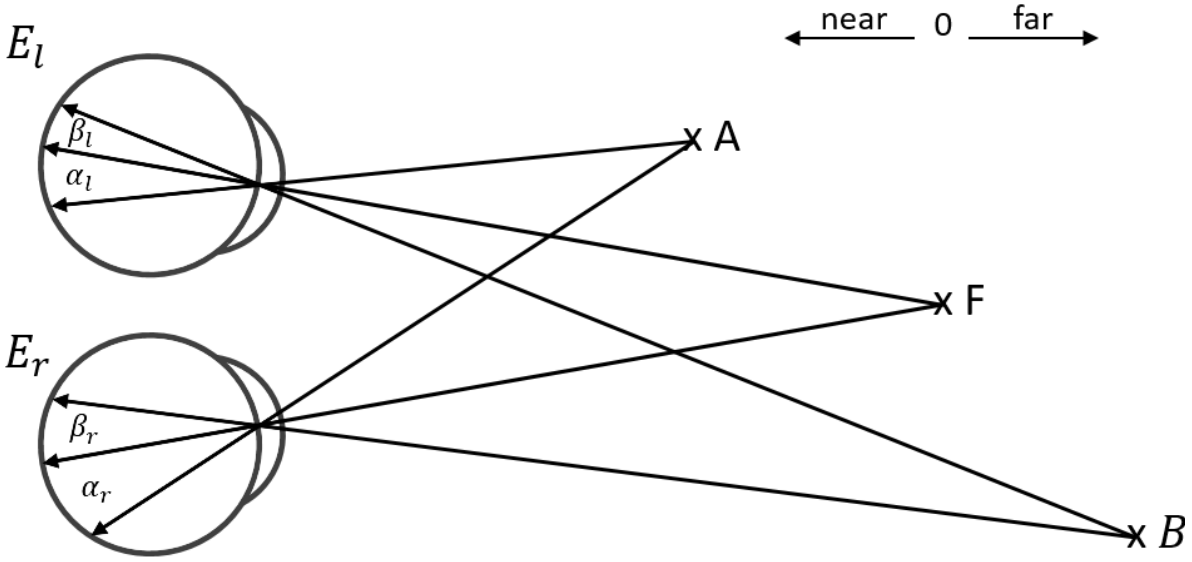


FIGURE 3-1

Disparity as an important cue in depth perception. Each single eye provides a single 2D projection of the 3D world (β, α), the projection of *F* is located at fovea (zero disparity). $\boldsymbol{\beta_l - \beta_r < \alpha_l - \alpha_r \rightarrow Depth_B > Depth_A}$

question is how to introduce neurophysiologically plausible stereopsis into technical systems by engineering underlying algorithmic principles of stereo-vision. The first attempt to answer this question was performed by David Marr who proposed a laminar network of sharply tuned disparity detector neurons, called Cooperative Network, to algorithmically model basic principles of the disparity detection mechanism of the brain [Marr 2010]. In 1989 Mahowald and Delbruck developed a micro circuit which was the first hardware implementation of Marr's cooperative network [Mahowarld & Delbrük 1989].

3.1.1 Correspondence Problem in Classical Vision

Besides the important role of 3D sensing in living systems, adding depth information into 2D visual information enables artificial systems, e.g. robots and assistive devices to operate in the environment with more reliability. Generally, in classic stereo vision, two cameras are used which are mounted on a common baseline to capture the scene from two different view-points. Geometrical characteristics of the stereo cameras can be formulated to map a single pixel of one image into a set of possible corresponding pixels in the other image [Hartly & Zisserman 2003]. Finding the matching objects or features in stereo images is called "correspondence problem". There are two general classic algorithms to solve the correspondence problem; area-based and feature-based matching. In area-based algorithms, usually the intensity of an area around a single pixel is

individually compared with a window around potential corresponding pixels in the other image. In feature-based matching, the correlation amongst features in each image is analyzed rather than intensity of pixels [Conradt et.al 2002]. Classical stereo-matching algorithms are computing-demanding, since they require processing a stream of frames that often contains redundant background information. This problem impedes applicability of classic algorithms in applications in which the processing time is crucial, e.g. Driving Assistive Devices and motion analysis [Ventroux et.al 2009].

3.2 Event Fusion vs Image Fusion, Stereoscopic Fusion in Silicon Retina

Dynamic Vision Sensors that mimic basic characteristics of human visual processing, have created a new paradigm in vision research [Lichtsteiner et.al 2008]. Similar to photoreceptors in the human retina, a single DVS pixel (receptor) can generate spikes (events) in response to a change of detected illumination. Events encode dynamic features of the scene, e.g. moving objects, using a spatiotemporal set of spikes (see FIGURE 3-1 (c)). Since DVS sensors drastically reduce redundant pixels (e.g. static background features) and encode objects in a frame-less fashion with high temporal resolution (about 1 us), it is well suited for fast motion analyses, tracking and surveillance [Conradt et.al 2009] [Drazen et.al 2011] [Müller & Conradt 2012] [Ni et.al 2012] [Osswald et.al 2017]. These Sensors are capable of operating in uncontrolled environments with varying lighting conditions because of their high dynamic range of operation (120dB).

Although DVS sensors offer some distinguished capabilities, developing event-based processing algorithms and particularly stereo matching, is considered as a big challenge in literature [Conradt et.al 2009] [Kogler et.al 2011] [Rogister et.al 2012] [Carneiro et.al 2013] [Camuñas-Mesa et.al 2014]. The fact that conventional frame-based visual processing algorithms cannot fully utilize main advantages of DVS necessitates developing efficient and sophisticated event-driven algorithms for DVS sensors. The main line of research in event-based stereo matching using DVS is focused on temporal matching [Kogler et.al 2011] [Rogister et.al 2012]. Kogler *et.al* proposed a purely event-driven matching using temporal correlation and polarity correlation of the events [Kogler et.al 2011]. Due to intrinsic jitter delay and latency in a pixel's response which varies pixel by pixel [Rogister et.al 2012], temporal coincidence alone is not reliable enough for event matching especially when the stimuli generate temporally-overlapping stream of events (i.e. when multiple different objects are moving in front of the cameras). Rogister *et.al* combined epipolar constraint with temporal matching and ordering constraints to eliminate mismatched events and have demonstrated that additional constraints can enhance the matching quality [Rogister et.al 2012]. Despite the fact that this method can partly deal with temporally-overlapping events, it still requires event-buffering for a time frame. To reduce ambiguity during the matching process, Carneiro *et.al* have shown that by adding additional cameras to the stereo setup (Trinocular vision vs. Binocular vision), it is possible to find unique corresponding matching event pairs using temporal and

epipolar constraints [Carneiro et.al 2013]. The results in this work show a significant enhancement in the quality of event-based 3D-reconstruction compared with other methods though Trinocular vision is not biologically realistic. Considering the epipolar constraint, it is necessary to calibrate cameras and to drive algebraic equations of 3D geometry [Hartly & Zisserman 2003]. Therefore, adding more cameras to the stereo setup will increase the complexity of the geometrical equations despite reducing ambiguity.

In this chapter a new fully event-driven stereoscopic fusion algorithm is proposed using Silicon Retinas. Event-driven matching means: as a single event occurs (caused by any motions or contrast changes in the scene), the algorithm should deal with it immediately and asynchronously without any need to collect events and construct a single frame. The main idea of our algorithm is borrowed from David Marr's cooperative computing approach [Marr 2010]. Marr's cooperative network can just operate on the static features to deal with the correspondence problem. In this work we have formulated a dynamic cooperative network in order to take into account temporal aspects of the stereo-events in addition to existing physical constraints such as cross-disparity uniqueness and within-disparity smoothness. The network's input includes the retinal location of a single event (pixel coordinates) and the time at which it has been detected. Then, according to the network's internal state (activity of the cells), which is shaped by previously fused events, disparity is extracted through a cooperative mechanism. The extracted disparity values can be further used for depth calculation of the events. The pattern of interaction amongst cells (suppression or excitation) applies physical and temporal constraints. In Section 3.4 we evaluated the proposed algorithm in several real-world experiments and the results demonstrate the accuracy of the network even with temporally-overlapping stimuli.

In the next section I will briefly describe the basic functionality of the Silicon Retina and in Section 3.3 the proposed event-based stereoscopic fusion method is elaborated in detail. In Section 3.4 experimental results are shown and finally, the conclusion and remarks are presented in chapter Section 3.5.

3.2.1.Neuromorphic Silicon Retina

In 1991 Mahowald and Mead developed the first silicon retina to bring principle functionality of the human retina into VLSI circuits [Mahowald & Mead 1991]. The basic operation of today's DVS sensors is similar to Mahowald and Mead's silicon retina whose pixels consist of a single CMOS photoreceptor to detect light intensity, differencing circuitry to compute change of the contrast or equivalently illumination, and comparator circuit to generate output spikes. FIGURE 3-1 (a) shows basic schematic of a single DVS pixel, where the light intensity is detected by a photoreceptor in the form of a current signal I and the current signal is amplified and transformed into a voltage signal V_p. The differencing circuit generates V_{diff} signal, which is proportional to change of log intensity ($V_{\text{diff}} \propto \Delta \ln(I)$). Finally, the comparator circuit compares the change of log intensity with preset thresholds. Therefore, if V_{diff} exceeds one of the ON or OFF thresholds (see FIGURE

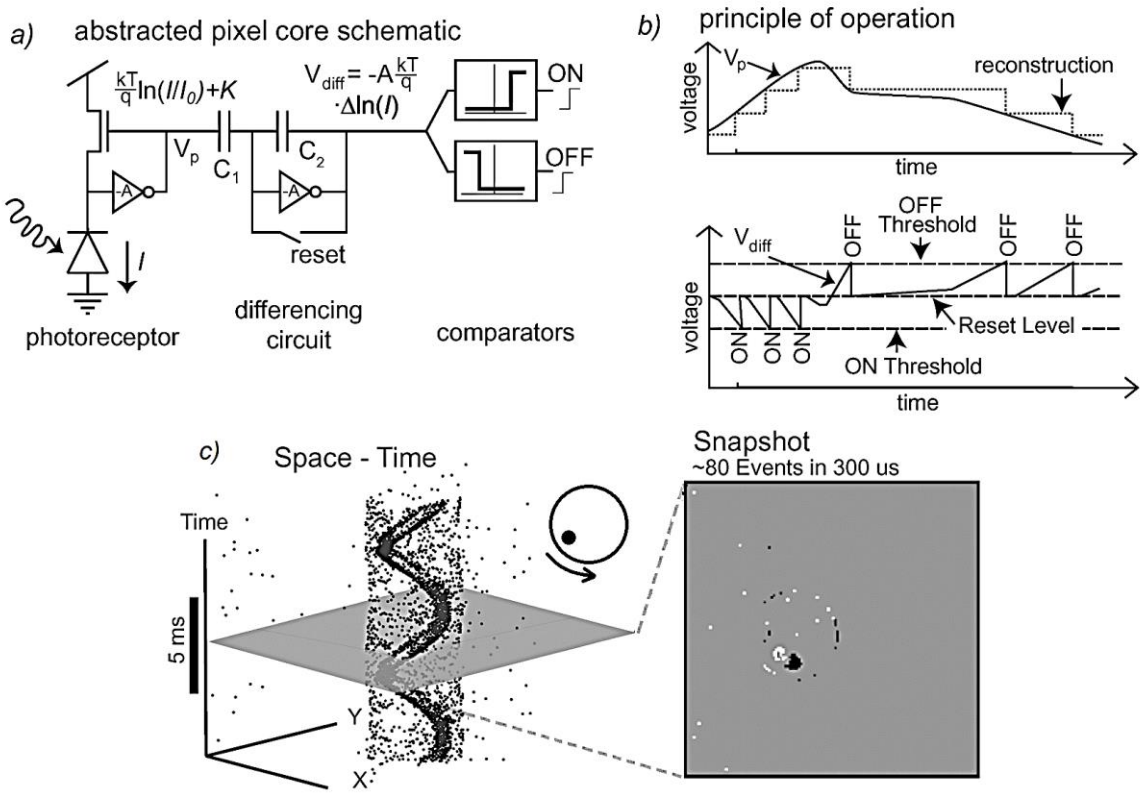


FIGURE 3-2

(a): abstracted circuitry of single pixel in DVS; (b): principle operation of single DVS pixel; (c) Space-Time representation of the event stream generated by a spinning disk, and single snapshot of the event streams in 300us, taken from [Lichtsteiner et.al 2008] with permission.

3-1 (b)), the sensor will signal out a single event. Each event consists of a data packet including pixel coordinates and the type of the event (ON and OFF events for negative and positive intensity change respectively). Finally, activated pixel will be reset into the base voltage (FIGURE 3-2 (b)). The encoding mechanism of the light using log intensity allows the sensor to operate in a wide range of illumination [Lichtsteiner et.al 2008]. FIGURE 3-2 (c) shows the Space-Time representation of an event stream for a rotating disk, and a single snapshot of the events within 0.3ms. The type of the events is indicated by dark and white pixels.

It is worth to notice that each pixel in DVS is independent from other pixels and the data communication is asynchronous. This means that events are transmitted only once after they occur without a fixed frame rate and are independent from each other. The sensor

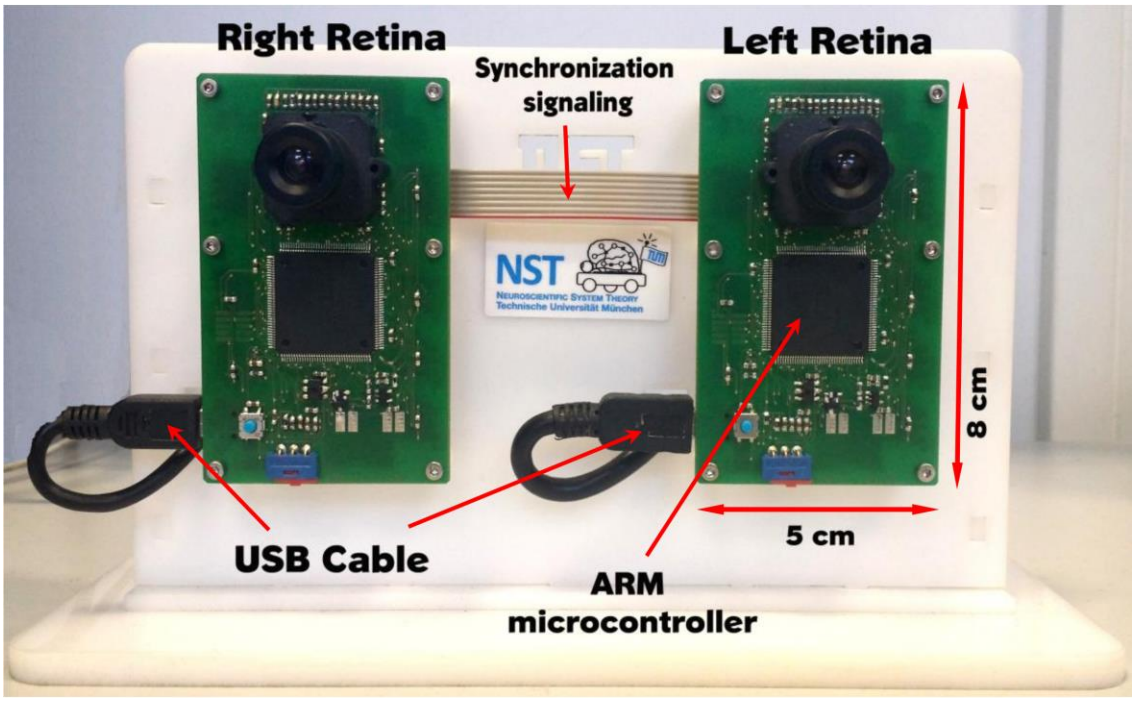


FIGURE 3-3
Synchronized embedded Stereo-DVS, eDVS4337.

chip that is used in this work is the *DVS128* sensor with *128×128* spatial resolution, *1us* temporal resolution and *120 dB* dynamic range of operation [Lichtsteiner et.al 2008].

3.2.2 Stereo Dynamic Vision Sensor

To fetch events and to stamp them with the time they have been generated, a supplementary circuit is required. In this work I use eDVS4337 circuit as a light, compact and low power solution that is used in several robotic applications and anomaly detection [Waniek et.al 2014] [Hoffmann et.al 2013] [Weikersdorfer et.al 2014] [Galluppi et.al 2014] [Conradt 2014] [Dikov et.al 2017]. This embedded system uses an LPC4337 ARM Cortex microcontroller to fetch events from the retina and to control the data path and communication links. The stereo setup is built using two eDVS mounted side-by-side so that silicon retinas are 10 cm apart (FIGURE 3-3). To synchronize two embedded boards, a Master/Slave signaling mechanism is performed on two microcontrollers before event fetching. Two separate USB links are connected to an external PC to read out event packets (FIGURE 3-3). Each packet consists of the retinal position of the event (x coordinate and y coordinate), the time-stamp t which is created by microcontrollers, and the type of the event which is called polarity p.

3.3 Principle of Cooperative Computation

Cooperative computing refers to the algorithms with distributed local computing elements that are interacting with each other using local operations. These operators apply specific constraints to the inputs in order to obtain a global organization and to solve a problem. The dynamics of these algorithms should reach a stable point given the inputs, to guarantee a unique solution for the problem they model. The pattern of interaction or equally the local operators, should be derived to computationally enforce the constraints amongst inputs. David Marr was the first who proposed a cooperative network to address the stereo matching problem. This network is composed of a matrix whose cells are created by the intersection of the pixel pairs in the left and the right images. Each single cell encodes the internal local belief in matching the associated pixel pairs. To derive the connectivity pattern between cells, two general physical constraints must be considered:

Cross-disparity uniqueness: reinforces a single pixel to possess one unique disparity value. Equivalently it means there must be a single unique corresponding pixel pair in each stereo image (in the case of no occlusion). So the cells that lie along the line-of-sight must inhibit each other to suppress false matching. For instance, in FIGURE 3-4 (a), given p_l on the left image as a retinal projection of the object P, there are two matches in the right retina, p_r or q_r. But, since just one of the candidates can be chosen, the cells that show the belief of p_r - p_l correspondence *i.e.* P in FIGURE 3-4 (a), should inhibit other cells that lie along disparity maps *i.e.* Q in FIGURE 3-4 (a).

Within-disparity continuity: Since physical objects are cohesive, the surface of an object is usually smooth and should be emerged by a smooth disparity map. Therefore, neighboring cells that are tuned for a single disparity or equivalently lie in a common disparity map, should potentiate each other to generate a spatially smooth disparity map (see FIGURE 3-4 (a)).

In spite of many unanswered questions about neurophysiological mechanisms of the disparity detection, nowadays it is widely accepted that mammalian brains utilize a competitive process over disparity sensitive populations of neurons, to encode and detect horizontal disparity [Zho & Quian 1996]. Similarly, in the cooperative network, the cells are sharply tuned for a single disparity value. Furthermore, the pattern of suppression and potentiation has implemented a competitive mechanism in order to remove false matching and to reach a global solution. Basically, the standard cooperative dynamics can extract spatial correlation of the static features in the scene, but the question arises how to formulate and to construct an event-driven cooperation process to deal with dynamic features, *e.g.* DVS event stream. In the following section I will address this question in detail.

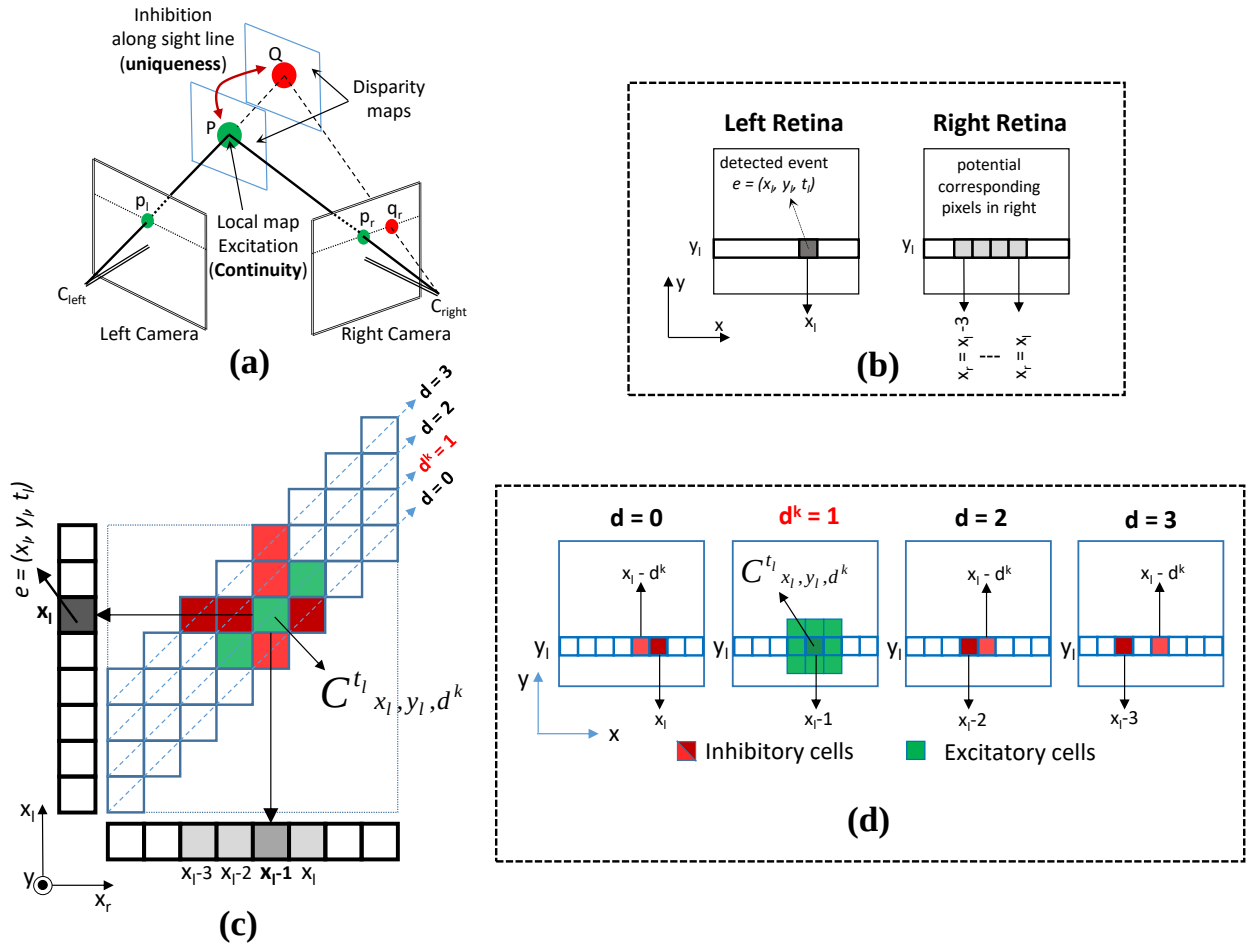


FIGURE 3-4

(a) Topological description of continuity and uniqueness constraints in stereopsis problem (b) Single detected event e= (xl, yl, tl) in left retina (dark grey) and potential corresponding candidates in right retina (light grey). (c) Cross-section scheme of the disparity maps, and stereo-retinas for y = yl (d) 2D disparity maps; the excitatory and inhibitory cells for Cxl,yl,dk are indicated by red and green respectively.

3.3.1 A Neural Model for Cooperative Event-based Fusion

Given event $e = (P_l, t_l)$ detected in the left retina at time t_l and $P_l = (x_l, y_l)$ as pixel coordinates (dark grey in FIGURE 3-4 (b)), the set of possible corresponding pixels in the right retina will be as follows:

$$S_e = \{(x_r, y_r) \mid x_l - d^{\max} \le x_r \le x_l, y_r = y_l\} \tag{3-1}$$

Where d^{max} is an algorithmic parameter that determines the maximum detectable disparity. For each possible matching pixel pair $P_l = (x_l, y_l)$ and $P_r = (x_r, y_r) \in S_e$, a single

cell $C_{xl,yl,dk}$ is created such that its temporal activity indicates the correspondence of that pixel pair. For instance, for the left event e in FIGURE 3-4 (b), cell $C_{xl,yl,dk}$ is created as the intersection of x_l in the left and $x_r = x_l - d^k$ in the right images. $d^k=1$ shows $C_{xl,yl,dk}$ is sensitive to disparity "*1*" (see FIGURE 3-3 (c)). So the network consists of a 3D matrix of the cells. The size of the matrix is $N \times N \times d^{max}$, in which N is the sensor resolution (FIGURE 3-3 (c)). If we expand the diagonal elements of the cooperative matrix into 2D maps, we can see a set of 2D disparity maps whose cells are specialized to detect one single disparity value (FIGURE 3-4 (d)). The cross section of the disparity maps for $y = y_l$ is shown in FIGURE 3-3 (c).

Since the stereo retinal images are aligned with each other and lie in a common baseline, in equation (3-1) we assume epipolar lines are parallel with y axis and potential corresponding pixels must have a common y coordinate. This assumption is true when stereo cameras are placed on a same surface and in parallel to each other [Hartly and Zisserman 2003]. To apply physical and temporal constraints on the input events, one should properly formulate the pattern of interaction among the cells. In order to support the continuity constraint and similar to the classic cooperative network, the neighboring cells lying on a common disparity map should potentiate each other creating a local pattern of excitation. The set of excitatory cells for a single cell $C_{x,y,dk}$ is indicated by green color in FIGURE 3-3 (c), (d) and can be described by following equation:

$$E(C_{x,y,d^k}) = \{C_{x',y',d^k} \,\|\, x - x' \le r, |y - y'| \le r\} \tag{3-2}$$

Having a unique possible matching pixel pair, the cells which lie along the line-of-sight should inhibit each other. So accordingly, there are two patterns of inhibition:

The first set of inhibitory cells which is shown in FIGURE 3-4 (d) by dark red, includes the cells that are topologically created by the intersection of the left pixel $P_l = (x_l, y_l)$, and all possible candidate pixels in the right ($P_r \in S_e$):

$$I_1(C_{x,y,d^k}) = \{C_{x',y',d} \,|\, 0 \le d \le d^{max}, d \ne d^k, x' = x - d, y' = y\} \tag{3-3}$$

A single candidate pixel in the right image (*e.g.* $x_r = x_l - 1$ in FIGURE 3-4 (c)), may have been chosen as a matching pixel for a former left event. Thus, a second set of inhibitory cells are selected in order to suppress this group of false matching (indicated by light red in FIGURE 3-4 (c), (d)).

$$I_2(C_{x,y,d^k}) = \{C_{x',y',d} \,|\, 0 \le d \le d^{max}, d \ne d^k, x' = x - d^k, y' = y\} \tag{3-4}$$

Descriptive features in DVS sensors are dynamic asynchronously-generated streams of events. Despite the fact that event matching based on exact temporal coincidence is not trustworthy, corresponding events are temporally close to each other. In other words, temporally close events have more probability to correspond to each other. Considering

TABLE 3-1

Flow of event-base cooperative stereo-matching algorithm

Algorithm.1, Event-driven cooperative stereo matching

Require: two synchronized retinas

Do for single event $e= (x, y, t)$

Construct set of possible corresponding candidates, $\mathbf{S_e}$ in equation (3-1).

for each corresponding pixel pair or equivalently $\{C_{x,y,dk} \mid 0 \leq d^k \leq d^{max}\}$

Find excitatory and inhibitory cells for $C_{x,y,dk}$, equations (3-2)-(3-4).

Compute activity of cooperative cell $C_{x,y,dk}$ according to equation (3-5).

End for

Do winner-take-all across all $C_{x,y,dk}$ cells.

If activity of winner cell is bigger than θ,

$D(e) = d^{WTA}$.

Else

Add small value ε to all corresponding cells activity.

End-If

Update the cells, (time and activity).

End Do

Wait for next event.

temporal correlation of the events, I have added internal dynamics into the cell activities such that each cell will preserve the last time it has been activated. Consequently, the contribution of each cell in the cooperative process can be weighted using a monotonically decreasing temporal kernel. From another point of view each cell keeps an internal dynamic by which its activity is fading over time like leaky neurons.

In consequence, the activity of each cell can be described by the following equations where W is temporal correlation kernel, E is the set of excitatory cells and I is the set of inhibitory cells for Cx,y,d^k, σ is a simple threshold function, α is a inhibition factor, and β tunes the slope of the temporal correlation kernel:

$$C^{t}_{x,y,d^k} = \sigma\left[\sum_{x'y'd' \in E} W^{t-t'}_{x'y'd'} C^{t'}_{x'y'd'} - \alpha \sum_{x'y'd' \in I} W^{t-t'}_{x'y'd'} C^{t'}_{x'y'd'}\right] \quad (3\text{-}5)$$

$$W^{\Delta t}_{x,y,d} = \frac{1}{1+\beta\Delta t} \quad (3\text{-}6)$$

Hess [Hess 2006] has analytically compared the *inverse linear* correlation kernel in equation (3-6) with *Gaussian* and *quadratic* kernels. He shows that this kernel can yield temporal correlation faster than Gaussian and quadratic functions without any obvious loss in quality.

Finally, the disparity for a single event $e = (x_l, y_l, t_l)$ can be identified by Winner-Take-All mechanism over activity of the candidate cells:

$$D(e) = \arg_{d^k} \max \left\{ C^{t_l}_{x_l, y_l, d^k} \middle| C^{t_l}_{x_l, y_l, d^k} \geq \theta \right\} \tag{3-7}$$

General flow of the algorithm is depicted in TABLE 3-1. The following parameters shape the algorithm and need to be tuned:

- Excitatory neighborhood, r in equation (3-2): tunes the smoothness of the disparity maps.
- Inhibitory factor, a in equation (3-5): tunes the strength of inhibition during cooperation.
- Activation function threshold, θ: each cell is active if integrated input activity in equation (3-5) becomes larger than the threshold.
- Slope of temporal correlation kernel, β in equation (3-6): this parameter can adjust the temporal sensitivity of the cells to input events. Larger factor means faster dynamics and sharper temporal sensitivity to the upcoming events.

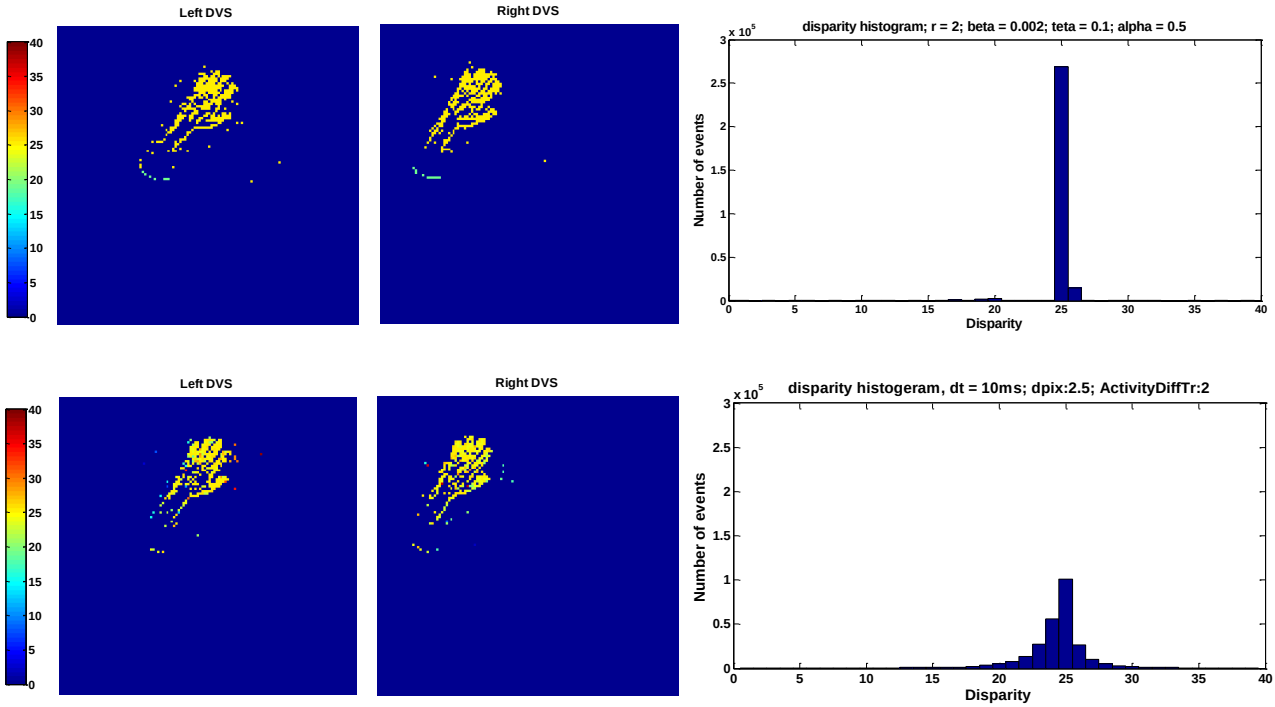
3.4 Experimental Results

The experimental stereo setup that is used in this work is described in chapter 2. The event packets are sent to a PC using two USB links, and the algorithm is implemented in MATLAB. There is no obvious standard benchmark in the literature to evaluate stereo matching algorithms using DVS. Rogister et.al used a moving pen as a simple stimulus which visually showed the coherency of the detected disparity in depth [Rogister et.al 2012]. To show the performance of the algorithm for temporally-overlapping stimuli, two simultaneously moving pens are used but the accuracy of the algorithm is not analytically reported [Rogister et.al 2012]. Kogler *et.al* have used a rotating disk (similar to FIGURE 3-2 (c)) as a stimulus to analyze the detection rate in an area-based, an event image-based, and a time-based algorithm [Kogler et.al 2011].

As the first experiment in this work, I create the disparity map of a single moving hand shaking in front of the retinas. In this experiment the algorithm has to deal with more complex stimuli than that of a single moving pen. The algorithm is executed without event buffering and the results for the stimulus located at 0.75m and 0.5m are shown in FIGURE 3-5 and FIGURE 3-6 respectively. For better visualization in these figures, a stream

of events is collected within 20ms and two stereo frames are constructed in the left and the right retinas. Then, the detected disparity for a single event is color-coded from blue to red for the disparity values varying from 0 to 40 pixels respectively. Moving the stimulus within two different known distances allows us to assess how coherent the detected disparity is with respect to the ground truth.

Since Rogister *et.al* have not quantitatively analyzed the performance of the algorithm in [Rogister et.al 2012], we have replicated this algorithm. The parameters of this algorithm for each experiment are analytically set to achieve best results. Single-shot extracted disparity maps using two algorithms are shown in FIGURE 3-5 and FIGURE 3-6 (for the stimulus placed at 0.75m and 0.5m respectively). As is depicted in the top row of FIGURE 3-5 (a) and FIGURE 3-6 (a), the extracted disparity maps using the cooperative network are perfectly distinguishable, and as the objects come closer to the sensors, disparity is increased. Although the algorithm proposed in [Rogister e.al 2012] is able to extract the disparity maps associated with the depths, the performance of this algorithm drops when the disparity is increased or equivalently stimuli come closer (compare



(a) (b)

FIGURE 3-5

(a): color-coded disparity map of a 20ms-long stream of events (in the left and the right retina) for a hand moving at 0.75m, (b): detection histogram within time of 5 sec.

Top row: the disparity maps and disparity histogram extracted by the cooperative network

Bottom: extracted disparity maps and disparity histogram using algorithm in [Rogister et.al 2012].

FIGURE 3-5 (a) and FIGURE 3-6 (a) bottom). Moreover, the disparity map extracted using the cooperative network is sharper around the ground truth, as compared with the algorithm proposed by Rogister *et.al* [Rogister e.al 2012]. In the top row of FIGURE 3-5 (a) and FIGURE 3-6 (a), the coherency of the disparity maps with ground truth values is clearly depicted. The smoother maps extracted by the cooperative network are intrinsically provided by the local pattern of excitation in equation (3-2).

Similar to the analysis performed in [Kogler et.al 2011], and in order to analytically evaluate the detection rate, I have created the disparity histogram using both algorithms for the events generated within a time period of 5 sec (FIGURE 3-5 (b) and FIGURE 36 (b)). The detection rate is the rate of the correct detected disparity with respect to the ground truth and is used as a performance criterion in the previous works [Kogler et.al 2011]. A range of detected disparity values within -1 and +1 of the ground truth value is considered as correct [Kogler et.al 2011].

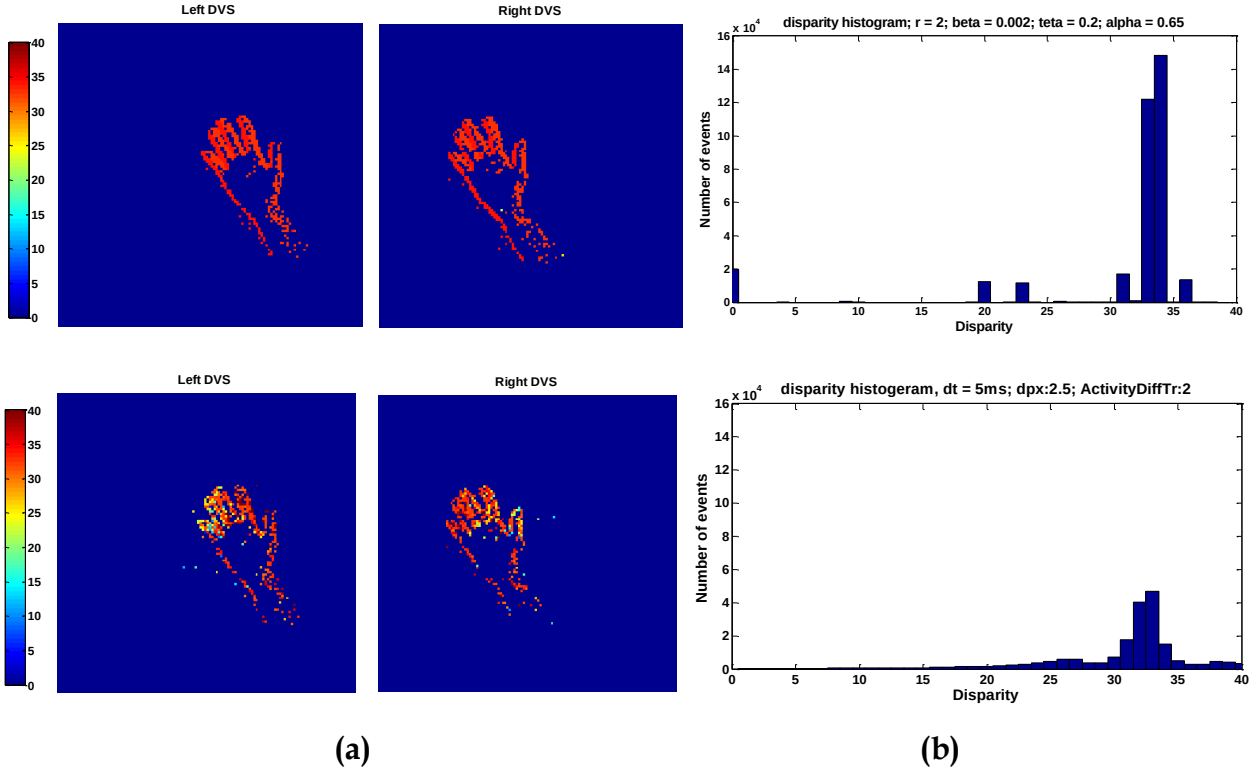


FIGURE 3-6

(a): color-coded disparity of a 20ms-long stream of events (in the left and the right retina) for a hand moving at 0.5m, **(b):** detection histogram within time of 5 sec.

Top row: the disparity maps and disparity histogram extracted by the cooperative network.

Bottom: extracted disparity maps and disparity histogram using algorithm in [Rogister et.al 2012].

It is illustrated in FIGURE 3-5 (b) and FIGURE 3-6 (b) that, when the cooperative network is used, only a small fraction of the events is mismatched. For the moving hand at 0.75m, 84% of the events are perfectly mapped onto the disparity map 25 or 24 (FIGURE 3-5 (b), top row), and for the stimulus located at 0.5m, 74% of the events are mapped onto the disparity maps 33 or 34 (FIGURE 3-6 (b), top row). These results show the advantages of the cooperative approach compared to the purely time-based event matching, in which the best average detection rate for a simple stimulus does not exceed 30% [Kogler et.al]. The average detection rates within 5 sec and using the algorithm in [Rogister et.al 2012] are 54% and 39% respectively for the stimulus placed at 0.75m and 0.5m (see the histograms at the FIGURE 3-5 (b) and FIGURE 3-6 (b), bottom).

To analyze the detection rate over time, the stream of events detected within 20ms-long time bins is collected, and the detection rate for each time bin is calculated. The graphs of the detection rate within a time duration of 10 sec for each experiment and using two algorithms are shown in FIGURE 3-7 and FIGURE 3-8. To compute detection rate in these graphs, the number of true matches divided by the number of whole events (including the events with unknown disparity) are calculated. For sparse time bins when the number of detected events has dropped, or equally when the stimulus is out of the overlapping retina's field of view, the momentary detection rate has dropped. This behavior is due the fact that when the stimulus is either partly located at the overlapping field of view, or it is out of the retina's field of view, many events are detected as unknown disparity and the detection rate significantly decreases. But, when the stimulus is located at both retina's field of views, the detection rate increases. The maximum detection rate of the algorithm proposed in [Rogister et.al 2012] does not exceed 70% for both experiments (red curve in FIGURE 3-7 top and FIGURE 3-8 top). Also, it is clearly shown that the detection rate of the cooperative network is always higher as compared to previous work particularly in the sparse time bins (FIGURE 3-8 top).

The results show that, the proposed network outperforms the algorithm proposed in [Rogister et.al 2012], in which an exact event-by-event matching is performed and the epipolar and the ordering constraints are used in addition to temporal matching to enhance matching. For each single detected event, previous algorithms will search for a corresponding event in the other image, whereas the proposed algorithm creates a set distributed maps of the cells, through which the cooperative computation is performed over the most recent detected events. The activity of a single cell indicates the internal belief of the network in a specific matching pair. Each single event inserts a tiny piece of information into the network such that the belief in the false matches are suppressed. Enhanced detection rate of the cooperative network compared with previous works, is due to the computational power of the event-fusion matching versus exact event-by-event matching.

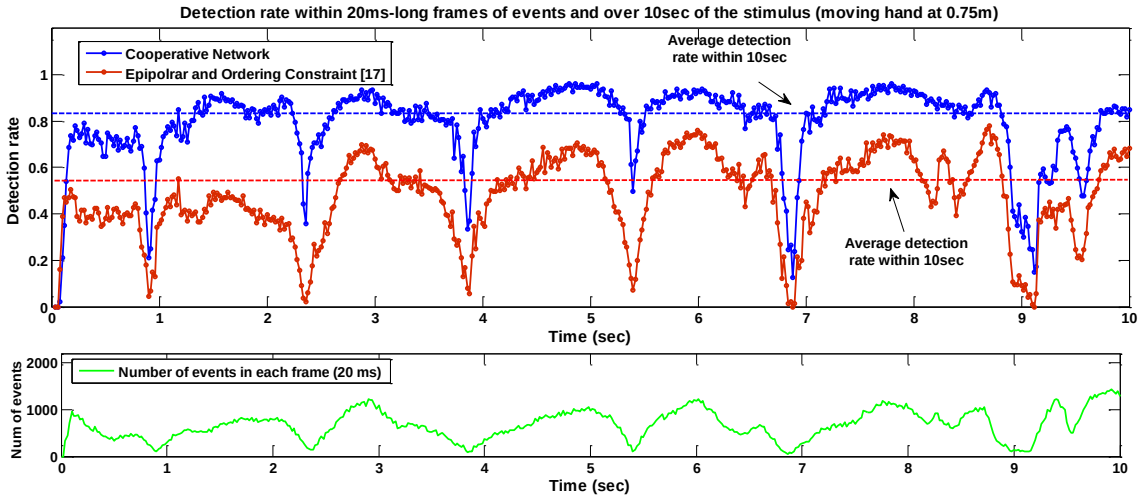


FIGURE 3-7

Top: Detection rate within 20ms-long time bins (frames) and over 10 sec of the stimulus (Moving hand at 0.75m). **Bottom:** number of events per time bin

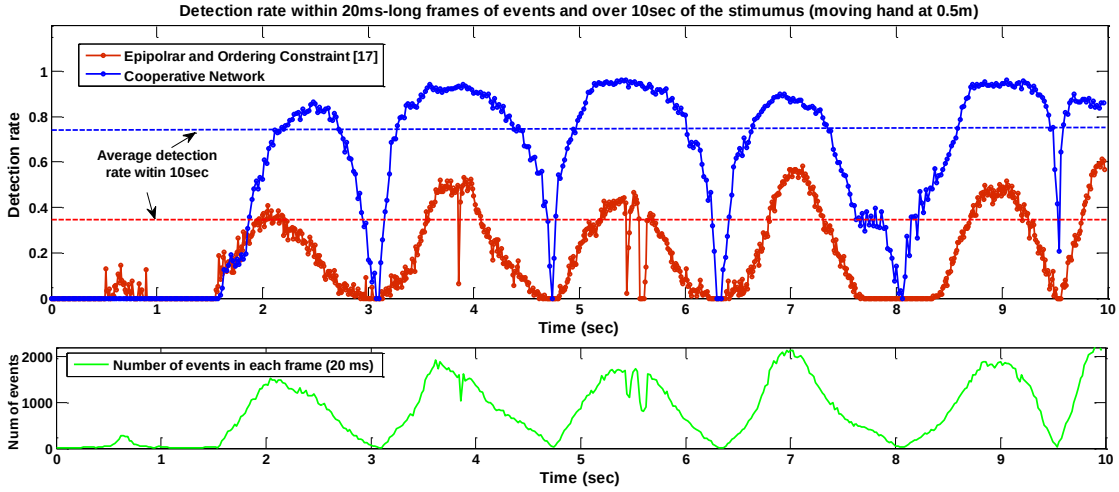


FIGURE 3-8

Top: Detection rate within 20ms-long time bins (frames) and over 10 sec of the stimulus (Moving hand at 0.5m), **Bottom:** number of events per time bin

Comparing the detection histograms in FIGURE 3-5 (b) and FIGURE 3-6 (b), the detection histograms for the stimulus located at 0.75m is sharper around the ground truth as compared with the stimulus placed at 0.5m. This shows there is more sensitivity of both algorithms to nearby objects and can be interpreted by the fact that in far distances objects often generate few events. Thus, the correspondence problem should deal with less ambiguity for the objects moving in far distances and it is easier to find matching pairs. This behavior has been observed in previous works [Kogler et.al 2011].

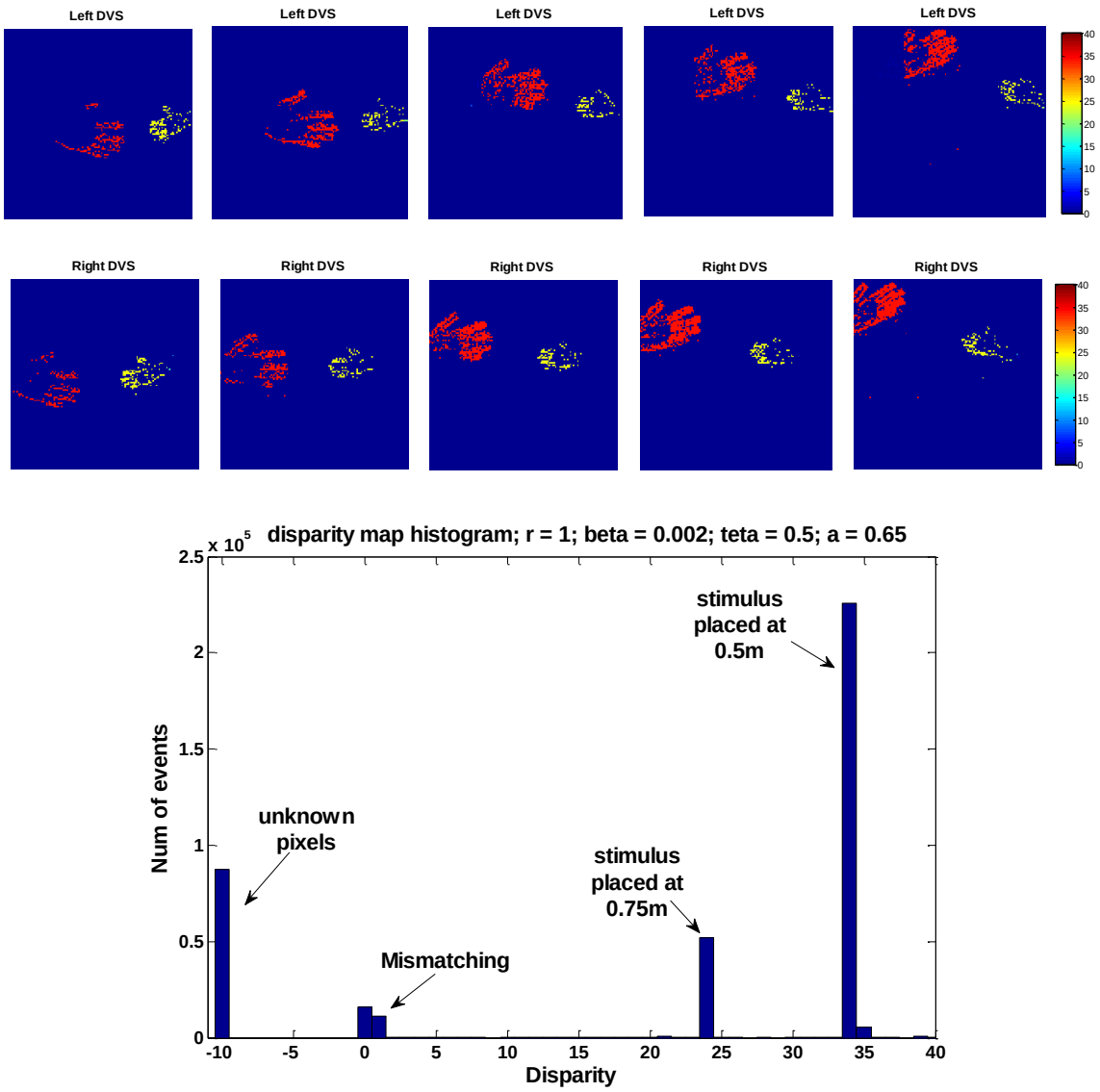


FIGURE 3-9

Top: color-coded extracted disparity maps over time for two moving hands (one at 0.75m and another at 0.5m). Each frame includes a stream of event generated within time of 20ms.
Bottom: Detection histogram for stream of events generated in 5 sec.

In order to evaluate the performance of the cooperative network in the cases with temporally-overlapping stimuli where the objects are located across common epipolar lines, we have created the disparity maps for two simultaneously moving hands, one is moving at 0.75m and another one is moving at 0.5m from the stereo DVS. In this scenario the algorithm should face considerably more ambiguity compared to the first experiment. The color-coded disparity values for a 20ms-long stream of events, and the detection histogram within 5 sec are presented in FIGURE 3-9. As is depicted in this figure, the

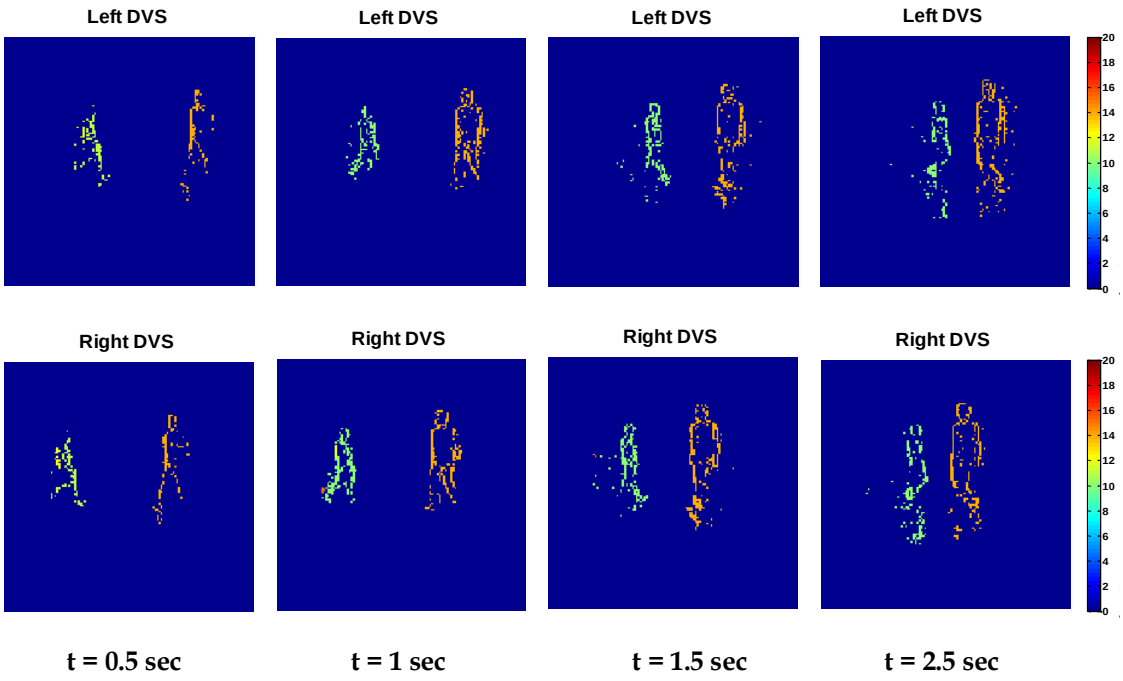


FIGURE 3-10

Color-coded extracted disparity maps over time period of 2.5 sec for two persons moving in the hallway with different distances from the stereo DVS. The events are collected within a time history of 25ms for each frame. **Green:** disparity = 10; **Orange:** disparity = 14

disparity maps which are purely computed in an event-based manner, is completely coherent with the depth of the moving objects (red color is corresponding to the events happened in 0.5m and yellow shows the events detected at 0.75m). In this experiment we have observed a considerable number of the events with unknown disparity (FIGURE 3-9 bottom). Unknown disparity happens when the activity of the winner cell in equation (3-7) does not exceed the threshold θ. The increased rate of the unknown disparity in the second experiment is a result of the increased number of the events that are out of the overlapping field of view.

Previous works required additional constraints, *e.g.* ordering constraint, orientation, pixel hit-rate, and *etc.*to reduce the increased ambiguity of the temporally-overlapping events [Rogister et.al 2012] [Camuñas-Mesa et.al 2014]. But, in the cooperative network, the second pattern of inhibition (equation (3-4)) suppresses a group of matching candidates that have been considered as corresponding pixels for a different object. This competitive process provides a mechanism to reduce false matches when multiple coincident clusters of events lie across or close to a common epipolar line and belong to different objects. In FIGURE 3-10 the extracted disparity maps for two moving persons in a hallway is presented in time. In this experiment most of the events have the risk of

multiple false matches, but the network can filter out a vast number of false events through the second pattern of inhibition.

The algorithm parameters for each experiment are listed in Table.2. Also, it is worth mentioning that the algorithm is implemented in 64-bit *MATLAB 2012b,* running on an *Intel Core i5 3.3GHz* processor with *16GB* RAM. The mean processing time per event with 25% CPU load (one processor is fully loaded) is listed in Table 2 for each experiment. One important aspect in most stereo matching solutions is the processing time. The average processing time in the proposed algorithm particularly depends on the number of disparity maps d^{max}. If we observe a single event as an elemental feature, average required processing time for a network with 100 disparity maps does not exceed *1.5ms* per event on our computing platform. Although achieved processing time might be still acceptable for some applications, natural parallelism of the cooperative network can speed up the processing on parallel hardware, which can be addressed in future research.

As a general rule of thumb, for sparse event-generating objects like the objects moving far away (or with slow motion), it is necessary to decrease the threshold of the cells activity θ (leads to more sensitivity of the cell), to allow sparse events to contribute in the cooperative process and not be cancelled as noise. Seemingly an adaptive homostacity mechanism [Remme et.al 2012] (*i.e.* adaptive θ) rather than global threshold setting, can help the network to detect sparse descriptive features. This work is worth to be investigated in future works.

3.5 Remarks

During the last decade, several attempts have been made to address the stereopsis problem using Neuromorphic Silicon Retina. Most of the existing stereo matching algorithms using DVS either are rooted in classical frame-based methods or temporal correlation. In order to fully take advantage of DVS sensors, developing efficient event-driven visual processing algorithms is necessary and remains an unsolved challenge. In this work I propose an asynchronous event-based stereo matching solution for Dynamic Vision Sensors. The main idea of the proposed algorithm is grounded in cooperative computing principle which was first proposed by Marr in the 80s. The classic cooperative approach for stereoscopic fusion operates on static features. The question that I have addressed in this chapter is how to formulate an event-driven cooperative process to deal with dynamic spatiotemporal descriptive features such as DVS events. To combine temporal correlation of the events with physical constraints, I have added a computationally simple internal dynamics into the network, such that each single cell can achieve temporal sensitivity to the events. Consequently, the cooperation amongst distributed dynamic cells can facilitate a mechanism to extract a global spatiotemporal correlation for input events.

Knowing the disparity of a moving object in the retina's field of view, I have used two basic experiments to analyze the accuracy and the detection rate of the proposed algorithm. Obtained disparity maps are smooth and coherent with the depth in which the stimuli are moving. The detection rate considerably outperforms previous works. In the second experiments the performance of the algorithm in response to temporally-overlapping events is evaluated. The results show that the cooperative dynamics intrinsically reduces the ambiguity of the correspondence problem when coincident cluster of events lie on the same epipolar lines.

Chapter 4

Proposed Neuro-Computational Models of Cue Integration

"In him there is no room for non-existence or imperfection"
— **Mulla Sadra** (1572 - 1640)

4.1 Motion-Cued Visual Attention using a Hierarchical Recurrent Neural Model

4.1.1 Introduction

Although we might not be aware of that, visual attention plays inevitable role whether directly or indirectly in perception, learning, and memory. Attention is the process of highlighting the relevant information and marginalizing the irrelevant signals out. Being relevant or not sometimes is determined by a top-down process. For example, when we are looking for our black shirt in the closet, we voluntarily focus on black clothes while we exclude other colors or probably objects. Sometime attention is driven by a salient sensory stimulus, e.g. the shattering sound or a bright object will draw our attention towards its location. The latter form of attention is called bottom-up attention [Bisley 2011]. There is a large body of research that significantly have contributed in understanding the mechanisms and neural correlations of visual attention in human (see these review papers: [Petersen & Posner 2012] [Bisley 2011]). Most of the computational models of visual attention are based on saliency-map [Filipe and Alexandre 2015] [Bruce & Tsotsos 2009] [Itti and Koch 2001] or a priority-map [Bisley 2011]; where the most salient stimulus is emerged through a winner-take-all competitive process. Rougier and colleagues have proposed a neural model of visual attention in which a dynamic interplay between inhibitory and excitatory synapses determines the location of the salient object [Rougier & Vitay, 2011] [Rougier & Vitay 2006] [Rougier 2006]. Here in this work it is argued that this approach can be scaled up for a top-down and voluntary attention scenario in which the symmetry of excitation and inhibition can be broken by a higher-order signal i.e. visual-motion, goal-associated signals. Studying the interplay

between perception and attention in multisensory research has been growing very recently [Emiliano 2012] [Rohe & Noppeney 2018]. The role of multisensory cue-integration in guiding attention allocation is theoretically investigated in this work. In this Chapter, I have proposed a hierarchical neural model in order to show how motion-cue can considerably enhance the quality of visual attention. The developed neural model is an extended version of that proposed by Rogier and colleagues proposed in [Rougier & Vitay 2006] and [Rougier & Vitay 2011]. The model that proposed by Rougier is based on Dynamic Neural Field [Amari 1977] and can describe some aspects of visual attention. However, in more complex scenarios for example when the focused object collides with a moving salient distractor, this model fails to register the location of the target. It is demonstrated that the hierarchical model proposed in this work can overcome this problem using a predictive mechanism.

In next section we will discuss about the role of hierarchical processing in attention allocation and attention control. In Section 4.1.2, we will describe the structure of the proposed neural model. In Section 4.1.3 the performance of this network using synthetic and realistic data is analyzed.

4.1.1.1. The Principle of Hierarchical Processing in Visual Attention

Hierarchical processing is a well-known and inevitable computational principle in Cortex which is directly involved in producing a wide range of cognitive functions [Felleman & Van Essen, 1991] [Riesenhuber & Paggio, 1999] [Cooper & Shallice, 2006] [Liu & Hou, 2013]. At each level of hierarchy, the information with specific level of complexity is preserved. The hierarchically registered information is reciprocally exchanged between cortical regions through feedforward and feedback projections. For instance, early visual cortices, i.e. *V1/V2*, collect sensory information from thalamus to create preliminary feature-maps within retinal-coordinate, e.g. spatial-map, spatial-frequency, retinal-disparity. Whereas, *V5* and *MST* regions that receive strong feedforward projections from *V1/V2*, consist of more complex neurons and compute visual-motion [Born & Bradley, 2005]. Thereafter more complex neurons in Parietal Cortex combine visual-motion information with signals from other modalities in order to form a more complex feature-map (spatial map) in body-centered and head-centered[42] coordinates [Sereno and Huang, 2014]. Posterior Parietal neurons receive strong feedback from Pre-Frontal Cortex. Both areas play a key role in sensory-motor tasks like saccade, visual tracking and smooth-pursuit [Uwe 2008], reaching [Vingerhoets, 2014], and particularly attention allocation [Saalmann et.al 2007]. Attention allocation is the process of selecting a location in the visual field that is behaviorally relevant and thereby is associated with a goal. The goal is most likely programmed in PFC and back-propagated

[42] For more detail see Chapter 2

to PPC in order to intervene action [Szczepanski et.al 2010]. More interestingly, Shomstein postulates that PPC is most likely the place within sensory cortex that bottom-up and top-down attention meet [Shomstein, 2012]. This region is known as one of the prominent multisensory convergence zones in sensory cortex and thereby can be possibly a place to investigate the twisted role of perception and attention.

One of the main advantages of such an anatomical and functional hierarchy is to have the allocated location of attention represented within different coordinates simultaneously, e.g. body-centered, head-centered, and eye-centered coordinates. It is evident that the feedback projections from parietal cortex to MT, and from MT to V1 moderate this process using gain modulation [Saalmann et.al 2007]. That means the activity of MT neurons whose receptive fields extend over the attended location is notably amplified. Looking for the underlying neural mechanism of attentional gain-modulation, Saalmann et.al has recorded the action potentials in lateral intra parietal, MT and V1 areas of macaques. When the monkey selectively focuses on the location of neurons' receptive field, the timing of activities become synchronized implying the fact that top-down feedback is used to propagate the allocated attention into early visual areas. More strikingly, Womelsdorf et.al observed that the receptive field of MT neurons in macaque is shrunk for those neurons that are tuned to the location of focused object. This can be explained by the gain-modulation driven by attention [Womelsdorf et.al 2008]. Following the principles discussed in this section, in the proposed hierarchical model MT neurons provide a modulatory feedback to the early visual areas or what is called in this chapter focus map. MT region is modeled by a motion sensitive population of neurons. These motion-detectors are laterally connected and receive information from a hidden-layer of neurons through a feedforward projection. Hidden layer consists of context neurons that preserve a history of the hidden neurons' activity. The hidden neurons are connected to the attention field using a feedforward connection. This 2-layered network is trained using Dynamic Error Back Propagation algorithm to give an estimate of the visual-motion for the attended object. Thereby the output of this neural layer modulates the attention field using a feedback projection. This feedback signal is in fact a prediction of the target's location in next time step. When the predictive neural activity overlaps with the sensory-driven neural activity in attention field, that would cause a stronger neural activity and thereby helps the observer to cancel out the colliding or salient distractors.

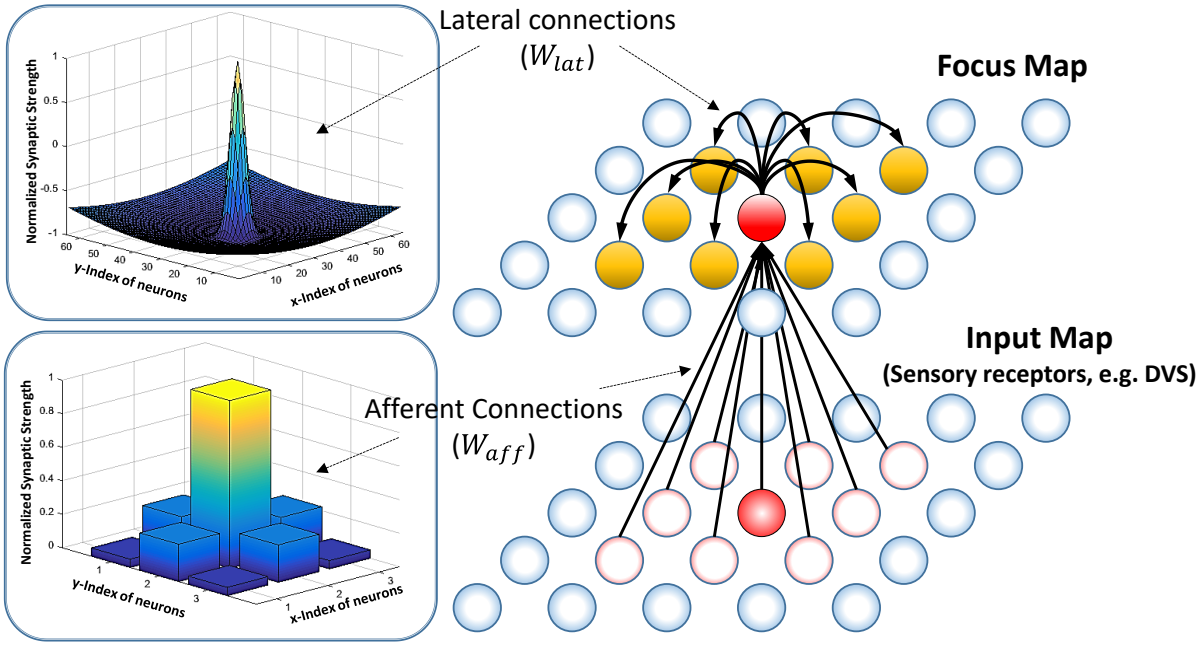


FIGURE 4-1

The basic Architecture of the Attention Network that consists of two neural fields: **Input-Map**, and **Focus-Map**. Input-Map encodes the activity of sensory receptors and it is connected to the Focus-Map neurons by afferent connection - W_{aff}; Focus-Map represents the location of the focused object in the field of view. Focus-Map Neurons are laterally connected according to W_{lat}. The activity of neurons are leaky which means it is exponentially decaying in time with a specific time-constant, which is the intrinsic characteristic of Dynamic Neural Field Model. In **left figures:** The pattern of normalized synaptic strength for lateral (top figure) and afferent (bottom figure) projections are plotted. As is depicted in bottom figure, the receptive field size for a single Focus-Map neuron is set to 3 pixels with 1 overlapping pixel.

4.1.2 Network Architecture

4.1.2.1 Focus-Layer, A basic Attention Network

The basic architecture of the motion-cued attention network is illustrated in FIGURE 4-1. The structure of this network is composed of two neural fields: Input-Map that represents the activity of Sensory Receptors (e.g. DVS events[43]), and Focus-Map that shows the location of the focused target in the retinal coordinate. Input-Map consists of $m \times m$ neurons to represent the relevant events that take place inside the field of view

[43] Dynamic Vision Sensor or Silicon Retina Technology. In Chapter 3, we have described the mechanism by which this new technology represents dynamic features into a stream of events (spikes). This asynchronous event-based representation of visual information, resembles the functionality of photoreceptors in human retina.

(events like moving objects, persons, cars, faces, red-colored objects, etc.). Similar to the overlapping receptive fields in retina and subcortical areas, the Input-Map is divided into $m' \times m'$ square patches with 1 radius equal to r_p. A single patch shares a set of overlapping pixels with neighboring patches and it is fully connected to a single Focus-Map neuron according to Equation (4-2). This kernel is called afferent connection and it is chosen to be Gaussian:

$$W_{aff}(X_i, X_f) = K_a e^{-\frac{\|X_i - X_f\|^2}{\sigma_a^2}}, for \ \|X_i - X_f\| \leq r_p \quad (4\text{-}1)$$

Where $K_a$is a factor that tunes the afferent synaptic strength; X_i and X_f are the retinal location that is encoded and preserved by the respective neuron of Input-Map, and Focus-Map; and σ_a tunes the width of Gaussian kernel W_{aff}. Note that, just those input neurons that lie inside the receptive field of a Focus-map neuron (the respective patch of input-map) have synaptic connections with that neuron. This data representation reduces the visual resolution and consequently leads to a significantly faster processing.

Similarly, Focus-Map is composed of $n \times n$ neurons that are laterally connected to each other. These neurons are segmented into $n' \times n'$ overlapping patches. The lateral connection follows a Mexican head function. This kernel (W_{lat} in FIGURE 4-1), enables a single Focus-Map neuron to excite its neighboring neurons while it inhibits the distant ones. This pattern of neural connectivity functions as a soft competitive winner-take-all mechanism [Rougier & Vitay 2006]. As a result, when the focused object emerged in the Focus-Map, it will cancel out the distractors while it will preserve the location of the target as a Gaussian of activity. Given X_f and $X_f{}'$ the encoding retinal location of a pair of neurons in the Focus-Map, the lateral weight is as follows:

$$W_{lat}(X_f, X_f{}') = K_b e^{-\frac{\|X_f - X_{f'}\|^2}{\sigma_b^2}} - K_c e^{-\frac{\|X_f - X_{f'}\|^2}{\sigma_c^2}}, K_b > K_c \ and \ \sigma_c > \sigma_b \quad (4\text{-}2)$$

Where $K_b$and K_c are the gain of excitation and inhibition patterns respectively, and σ_c and σ_b are the standard deviation of Gaussian profiles for inhibitory and excitatory synapses.

Having the neural connectivity and general structure formulated, the dynamics of the neural activity is modeled by the following Equation. This model of neural computation is called Dynamic Neural Field [Sandamirskaya 2014]:

$$\frac{\partial u(X_f,t)}{\partial t}\tau = -u(X_f, t) + \sum_i W_{aff}^{(X_i,X_f)} I(X_i, t) + \sum_{f'} W_{lat}^{(X_f,X_{f'})} F[u(X_{f'}, t)] + gM(\mathrm{u,t}) \quad (4\text{-}3)$$

$$F(u) = \frac{1}{1+e^{-(u-\beta)/\gamma}} \quad (4\text{-}4)$$

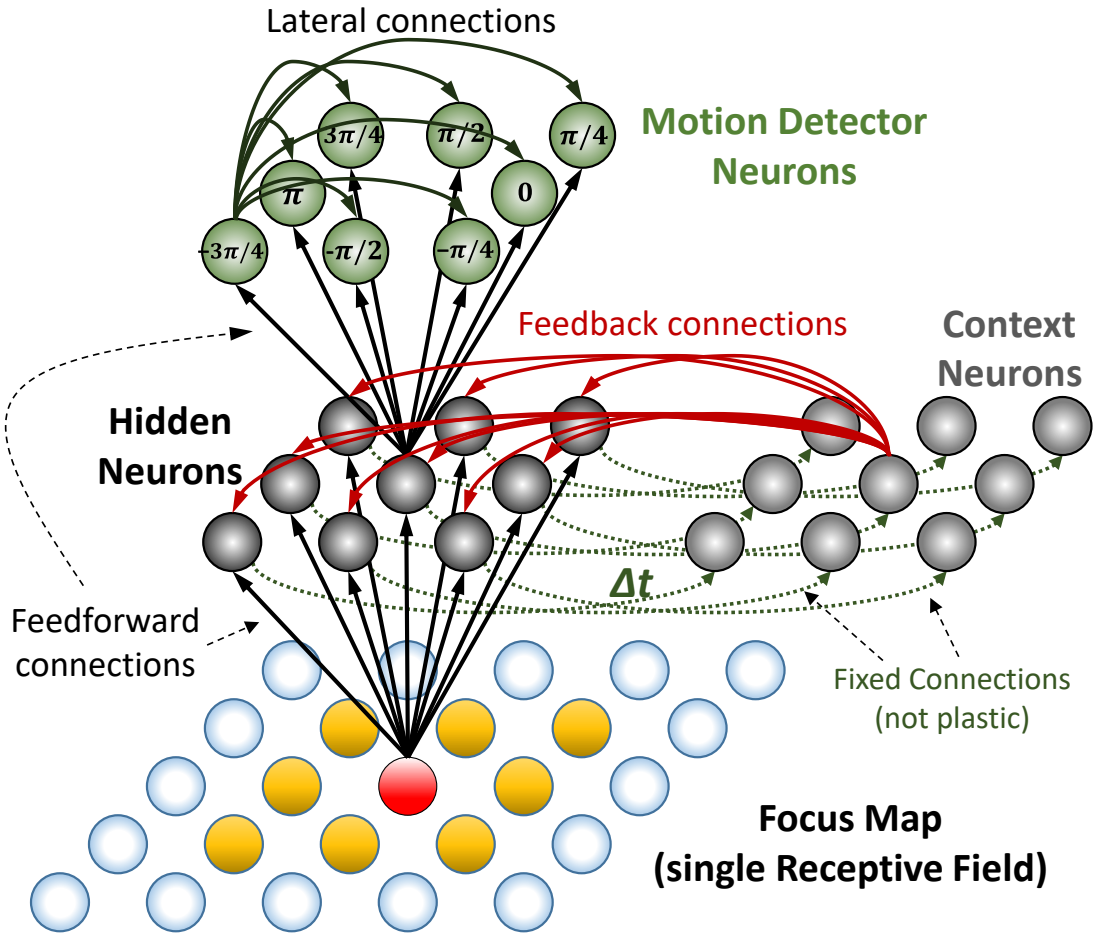


FIGURE 4-2

The architecture of the motion estimation network. The functionality of this network is to estimate the motion direction of the focused target which is captured by Focus-Map. Similar to functional organization of MT neurons, for each single patch of Focus-map, a ring of 8 motion-sensitive neurons are defined to encode the estimated direction. Each patch resembles the visual receptive field of motion detector neurons. Focus-Map neurons are fully connected to a pool of *p* hidden neurons, while the hidden neurons are also fully connected to each motion-detector neurons. A history of neural activity of hidden neurons is preserved within context neurons (in fact they are a delayed copy of hidden neurons). These neurons are connected to hidden neurons through a feedback connection.

In Equation (4-3), $u(X_f, t)$ represents the level of activity[44] of Focus-Map neuron X_f. In other words, this 2D variable shows the likelihood of the target's location at each time. In (4-4) τ represents the time constant of neural dynamics, or equivalently the leakage of membrane potential; I is the input current form input-map or equivalently the level of activity of i^{th} input neuron, and $F(\cdot)$ in (4-4) is the Activation function which is usually

[44] Sometimes referred as membrane potential.

linear-threshold or sigmoid [Sandamirskaya 2014]. And finally, $M(\cdot)$ is an additive neural activity which is emerged by recurrent connections from motion sensitive layer to Focus-Map. This term provides a new evidence regarding the possible location of the target in next time, given the previous neural activity of Focus-Map. From another point of view, $M(\cdot)$ implements an additive attentional gain modulation. Attentional neural modulation can be both additive and multiplicative, depending on the size of stimulus and attended field [Reynolds & Heeger 2009]. In general, whenever the size of stimulus is smaller than the attention field, an additive gain modulation is observed [Reynolds & Heeger 2009]. If the size of stimulus becomes comparable to or greater than attended field, the neuron shows a multiplicative modulation. In the next section, we will elaborate how this modulatory neural activity is computed and why it functionally makes sense.

4.1.2.2 Motion Sensitive Layer

As we discussed in Section 2.3.4, given the current location of the focused object, e.g. a prey, and its velocity, the nervous system is able to predict the next state of the target before providing a new sensory evidence (see FIGURE 2-10). To incorporate this a-priori information in the neural model and in the context of visual attention, we have added a motion-sensitive neural network that takes the momentary activity of the Focus-Map as input and estimates the direction of motion for the captured target. The general architecture of this network is depicted in FIGURE 4-2. Middle-Temporal area of primates Visual Cortex plays an inevitable role in representation and coding of visual motion [Britten, 2003]. Synonymous to MT neurons that are tuned specifically for a direction of motion and for specific area of visual field, for each single patch of Focus-Map, we have defined a ring-population of laterally connected neurons (motion-detectors in FIGURE 4-2) [Born & Bradley 2005]. The ring-population is composed of 8 neurons, each is tuned for one of the 8 possible directions around the patch. Therefore, they are tuned to a range of angles from 0° to 315° with a resolution of 45°. This structure accommodates a single portion of visual field with a dedicated population of motion neurons, like the way visual cortex implements this functionality. However, as compared with MT neurons, motion-detectors in our model is simplified so that the neurons are specifically sensitive to one of 8 possible directions of motion, and a reasonable range of velocity. If the target passes through a single patch, the respective motion estimation network will be activated in order to estimate the direction of motion, and thereby to apply the predictive neural activity i.e. $M(\cdot)$ to Focus-Map. Each single patch of the Focus-Map is fully connected to a pool of q hidden neurons within a feedforward connection. To preserve the state of the real world in the previous time, the neural activity of hidden neurons, delayed by Δt, is preserved within another pool of q neurons called context neuros. Context neurons are also fully connected to hidden neurons through a feedback connection (red connections in FIGURE 4-2). Finally, the hidden neurons pass the superposition of information regarding the state of the target in present time (preserved in hidden neurons) and

previous time (represented by context neurons), to the motion-detectors through a feedforward connection. Motion-detectors are also laterally connected using a Mexican-hat kernel (formulated in Equation (4-2)). From computational point of view, this lateral connection in fact performs a smoothing and de-noising process on final estimate which results in a more stable locating and tracking the target [Born and Bradley 2005]. The activation function for output and context neurons are liner, while the hidden neurons are sigmoid.

Despite the fact that MT neurons are tuned to a specific range of speed along with the direction of motion [Krekelberg et.al (2006)], we have not included speed-sensitivity in our model for the sake of simplicity. The architecture of the motion estimation network is borrowed from Elman Recurrent Neural Networks which is developed in 90s to deal with sequence-prediction problem and dynamic system identification [Zimmermann and Neuneier 2000] [Elman 1990]. These tasks are beyond the power of a multilayer perceptron network. In Section 4.2.1.3 we will discuss in detail how to generate the proper input-output features to feed into this network, and thereby how to train the network.

It is important to note that the size of the patches in Focus-Map determines the size of receptive fields for motion-detector neuros. This parameter must be tuned large enough to capture the direction of motion for the target. Given a Gaussian-like afferent and lateral connections in (4-2), the outcome of the Equation (4-3) for almost any arbitrary input I, is a Gaussian-like bump of activity in focus-map. The size of this emerged bump of activity is proportional to the width of lateral excitatory connection, i.e. σ_b^2. Consequently, and as a rule of thumb, the size of a Focus-Map patch must be at least twice bigger than the diameter of the emerged bump. On the other hand, neural dynamics time constant τ should be small enough compared with the velocity of target. Otherwise the Focus-Map cannot smoothly follow the trajectory of focused object, or the bump would abruptly jump to a new place.

So far, we have described the structure of the motion estimation network that takes the emerged bump of activity in Focus-Map as input, in order to estimate the direction of motion, and thereby to generate $M(\cdot)$ in (4-3). But, the question is how the motion information can be fused into a neural field which represents location information, i.e. $u(x,t)$ in (4-3). Having the direction of motion estimated and preserved within the activity of motion detectors, and given the most active patch[45] at hand, the probability of the target's location in the next time step can be modeled by a 2D uniform distribution extends over the area towards which the target moves. Note that we have not included the velocity in the model. Therefore, the next location of the target cannot be specifically determined by adding a factor of velocity vector to current location. Instead, it is modeled

[45] Active patch is a patch in the Focus-Map with highest neural activity. Equivalently, the active patch is supposed to capture the focused object.

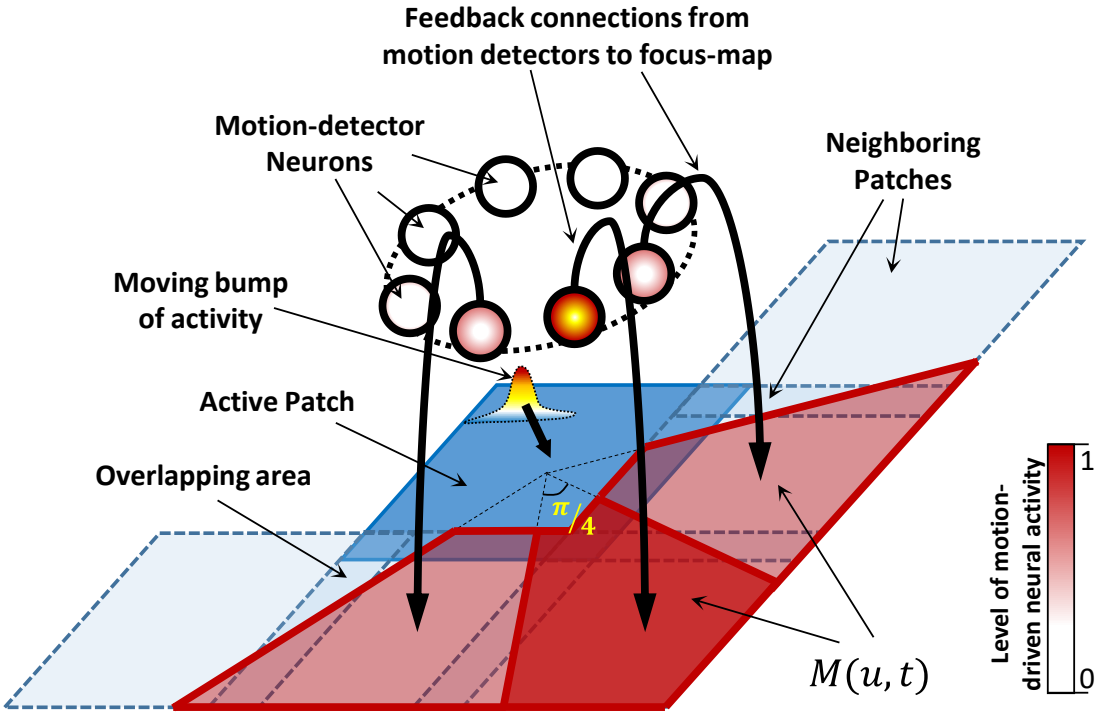


FIGURE 4-3

The basic pattern of feedback connections from motion sensitive layer to the Focus-Map is illustrated. For a given active patch where the level of neural activity exceeds a specific threshold and the bump of activity is moving across that patch, the motion-estimation network will be activated to generate the corresponding predictive neural activity - $M(u,t)$. A specific area of neighboring patches should be activated according to the neural activity of respective motion-detector neurons. The color-coded areas in Focus-Map represent the pattern of feedback connections from motion-layer to focus-layer and that also includes the overlapping areas.

by a set of locations chosen crosswise with respect to direction of motion, and with identical probability of occurrence. This uniformly distributed area of possible locations in visual field, is in fact $M(\cdot)$ in (4-3). In addition to direction of motion, the resolution of estimation also shapes this area. For example, if the target moves from left to right, at the next time step it is more probable to locate it somewhere between active patch and the neighboring patch at right-side. Moreover, since the resolution of estimation is 45°, just a sector of neighboring patch must be exclusively activated. In FIGURE 4-3, it is depicted how to incorporate these constraints and assumptions into the model. The pattern of feedback projection from the motion-sensitive layer to the focus layer is demonstrated by distinct sectors colored according to the level of activity of the relative motion neurons.

4.1.2.3 Training Motion Sensitive Layer

The afferent and lateral connections of focus-layer are both Gaussian (see FIGURE 4-1). As a result, the location of the focused object will be emerged in the Focus-Map as a Gaussian bump of activity. The size of this bump is proportional to σ_b^2 (width of excitatory region). Therefore, to train the motion estimation network, the Input-Map is fed by a hypothetical target moving toward the desired angle of motion so that a bump of activity in the Focus-Map is emerged (see FIGURE 4-4 left column). The activity of Focus-Map then, is used as input signal to the network (see FIGURE 4-4 middle column). The desired angle is in fact the analog training signal which determines the activity of each motion-detector neuron. The encoding mechanism, i.e. coding an analog signal through the activity of a population of neurons, is governed by a clamped cosine function. This encoding scheme is formulated in the following Equation in which θ_d is the desired angle of motion, v_j is the neural activity of j^{th} motion-detector neuron, and θ_j is its preferred angle. Preferred angle for a single motion-detector is the analog value of a stimulus at which the neuron exhibits maximum activity.

$$v_j(\theta_d) = \begin{cases} \cos[(\theta_j - \theta_d)], & \cos[(\theta_j - \theta_d)] > 0 \\ 0, & \cos[(\theta_j - \theta_d)] \leq 0 \end{cases} \tag{4-5}$$

There are two reasons behind choosing a clamped-cosine function for encoding. First it enables the motion-sensitive layer to exclusively excite the area of the Focus-Map that extends over the range of $[\theta_d - \frac{3\pi}{8}, \theta_d + \frac{3\pi}{8}]$. This area excludes the orthogonal directions, i.e. $\theta_d + \frac{\pi}{2}$ and $\theta_d - \frac{\pi}{2}$, to be activated. Secondly, this kernel is Gaussian-like and resembles the tuning-function of MT neurons. The structure of the motion estimation network is identical for all patches. So, it is required to train a single network and once it is trained, that can be replicated for all patches with identical parameters. We have generated 15878 pairs of input-output data from which 11734 points are used to train the network, and 4144 points are used to test the performance of the classifier at each iteration. At each trial, a Gaussian bump of activity moves across a single patch towards one of the eight possible directions. The point from which the bump enters the patch is also set so that all possible scenarios are included. Even though the velocity is constant and set to 1 pixel/frame, it is practically possible to generate training data from bumps moving with a range of different velocities. In next section we will show that the trained network is still able to estimate the direction of motion for targets moving twice faster and slower than that of training patterns. The general parameters of the network are listed in TABLE 4.1

Having the set of input-output features determined, i.e. $\left\{(U_j, V_j^d) \mid j = 1,2,3,\dots,p;\ U_j = \{u(x_f,t)|x_f \in active\ patch\}\ and\ V_j^d = \{v^k(\theta_d)|d = 1,2,\dots,8\}\right\}$, we can train the synaptic weights using Dynamic Error-Back-Propagation algorithm [Pham & Liu, 1996]. At each

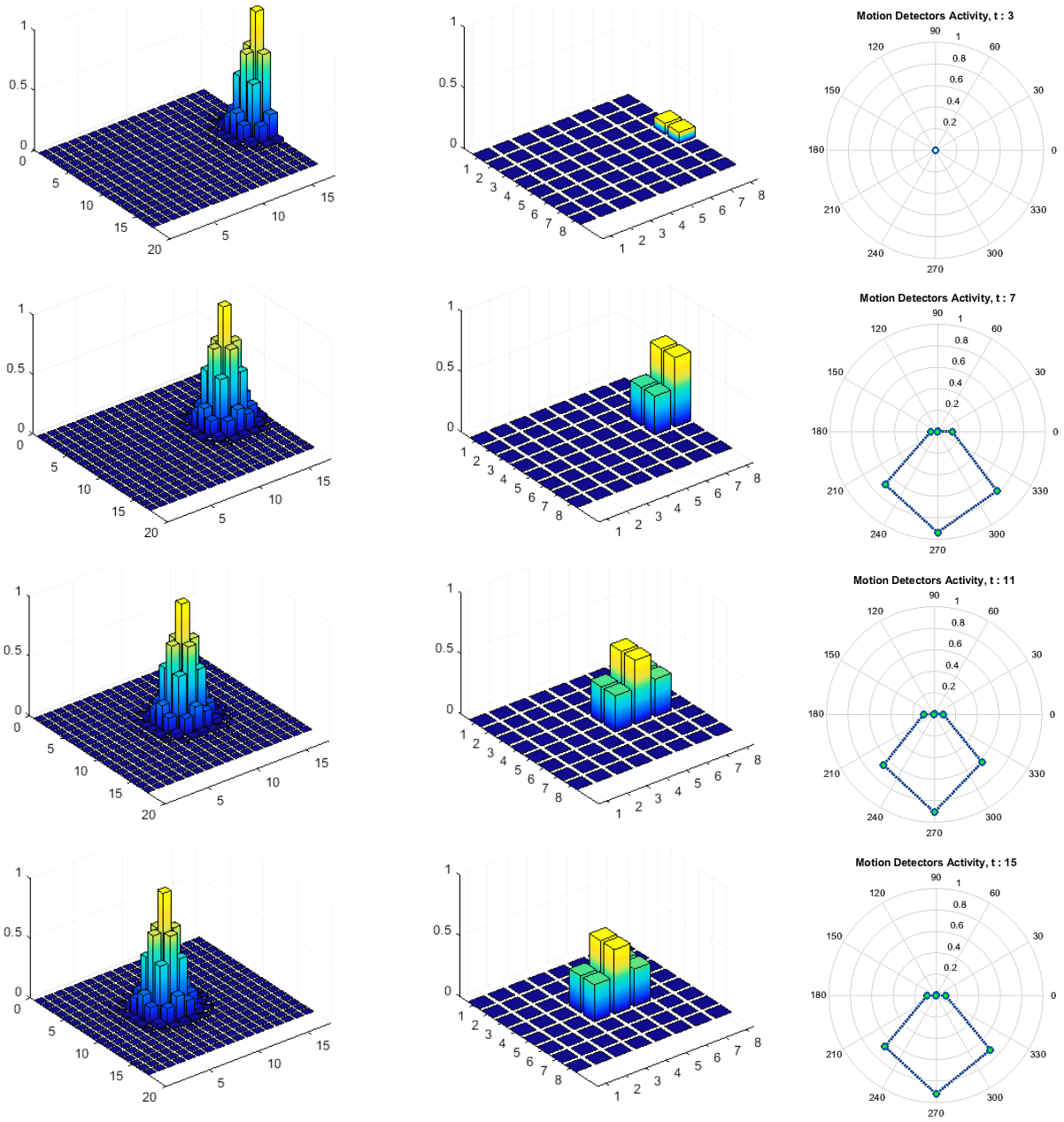


FIGURE 4-4

The performance of motion estimation network after training is illustrated within an example. In the **left column**, a Gaussian bump of activity moves across a segment of the Input-Map whose neurons are projected to a single 8×8 Focus-Map. In the **middle column**, the emerged bumps in the Focus-Map are shown. And in the **right column**, the output of motion estimation network is plotted within polar coordinate where the activity of a single motion detector neuron is represented by a single point. The angle shows the preferred direction of motion for a single motion detector, while ρ represents its level of activity. The direction of motion for this example is set to 270°. As is demonstrated in these figures, the networks is able to estimate direction of motion correctly. It is noticeable that the network will not respond to an inactive patch of Focus-Map (top figure). Each row corresponds to the activity of neurons in time.

TABLE 4-1 Parameters of the network

K_a	σ_a	K_b	σ_b	K_c	σ_c	τ	m	m'	n	n'	p	g
0.75	0.03	200	0.3	0.6	0.9	0.6	128	3	64	8	32	*0.4*

iteration of EBP algorithm, first the square error between the desired output V_j^d, and output of the network fed by U_j is computed. Then, the weights are updated by a negative factor of Error-Gradient. This modification scheme is called gradient descent method and it is formulated in (4-6). In this Equation, V_j is the output of the network evaluated by U_j as the input; E_j is error function, w is a vector of synaptic weights, and α is the earning rate:

$$\Delta w = -\alpha \frac{\partial E_j}{\partial w}, E_j = \frac{1}{2}\left(V_j^d - V_j\right)^2, 0 < \alpha < 1 \quad (4\text{-}6)$$

Dynamic EBP is a modified version of EBP in which the dependency of context neurons' activity (see FIGURE 4-2) to the activity of hidden neurons in the previous time step, is taken into account. This leads to a modified error function by which the learning process will be more stable [Pham & Liu 1996]. In FIGURE 4-5 the evolution of weights in favor of minimizing MSE for training data and test data set is depicted. As we can see in this figure, Root Mean Square Error is saturated after about 5000 iterations of the algorithm. This means that the synaptic weights are updated enough to model the desired functionality, i.e. generating the desired output given the input pattern. As another notable fact, RMSE for the test data is slightly greater than that of the training data. After 10000 iterations of Equation (4-6), RMSE for the test data set is equal to 0.0023, and for the training data set it is equal to 0.0016. In FIGURE 4-4 right column, the output of the motion estimation network in response to a test pattern is plotted in polar coordinate. As is depicted, the motion detectors remain silent if the activity in the patch is less than a specific threshold. When the activity exceeds the threshold, the network will estimate the angle of motion (FIGURE 4-4 right column). It is also worth to note that the lateral connection between motion-detectors are removed during learning.

4.1.3 Performance Evaluation and Results

4.1.3.1 Noise Sensitivity Analysis

To evaluate the sensitivity of the model to noise, we have simulated a similar experiment that Rougier and Vitay performed [Rougier & Vitay 2006]. A hypothetical ball is used as a synthetic target that moves along a circular path with radius r=32. Once the target is captured in the Focus-Map, i.e. after 10 frames, the activity of Input-Map neurons including those representing the target's location, are perturbed by additive white noise. Equation (4-7) formalizes this stimulus:

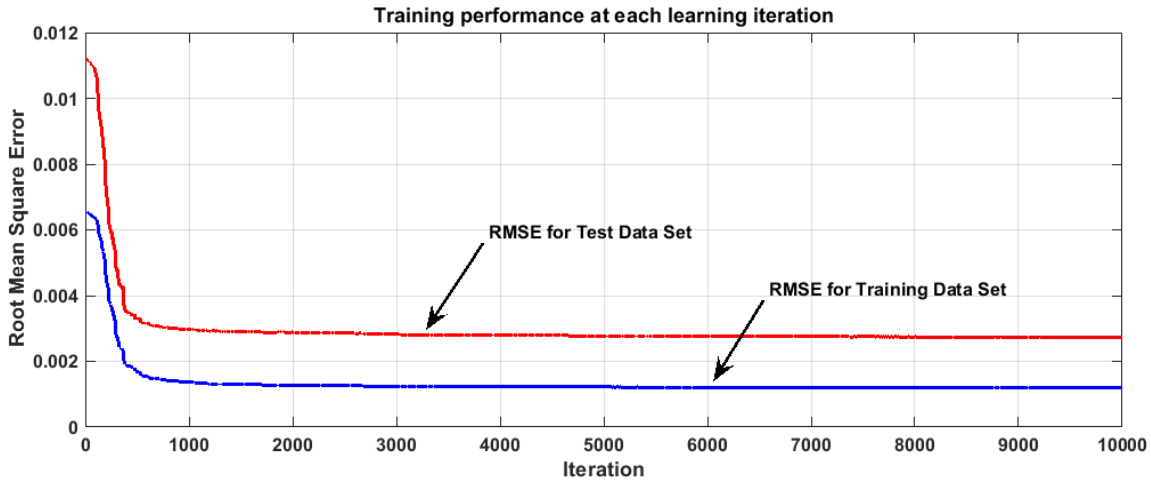


FIGURE 4-5

The diagram shows the evolution of the synaptic weights in the motion estimation network. The Root Mean Square Error for the test data and training data is separately plotted at each iteration of the learning algorithm. As we expect the error for test data is greater than training data. For both test-set and training-set, the profile of RMSE is saturated after 5000 epochs. The stop condition is to check whether the RMSE exceeds a preset value, or the algorithm runs for a specific number of iterations. After 10000 iterations, the final MSE for the test set and training set is 0.0023 and 0.0016 respectively.

$$I(X_i) = I_p \frac{\|X_i - X_c\|^2}{w^2} + n_\beta; X_c(t) = \begin{bmatrix} rcos(\theta(t)) + x_0 \\ rcos(\theta(t)) + y_0 \end{bmatrix} \quad (4\text{-}7)$$

In this Equation, $I(X_i)$ represents the activity of the input neuron located at X_i while I_p is the maximum neural activity in the input layer; X_c is the location of ball, w tunes the size of the ball, r is the radius of circular path (see FIGURE 4-6 (a)). The noise matrix is denoted by $n_\beta \in R^{n \times n}$ whose elements are statistically independent and identically distributed. In FIGURE 4-6 (b) a single frame of Input-Map which is polluted by noise is depicted (noise intensity is set to 50% in this example). The respective response of the Focus-Map neurons to the input is shown in FIGURE 4-6 (c). Finally, (x_0, y_0) in (4-7) is the point from which the target starts to move and for this experiment it is set to fovea, i.e. $(64, 64)$. The velocity of the ball is assumed to be constant and is set to 1 pixel/frame. Noise intensity is noted by a subscript β which in fact is noise variance. The parameters of this network are listed in TABLE 4-1.

Similar to the performance criteria used by [Rougier & Vitay 2006], we have measured the distance of emerged bump of activity in the ocus-Map to the original location of target in the Input-Map. In better words, the focused location is determined by weighted averaging over the activity of the Focus-Map neurons. Given the stimulus of Equation (4-7), the measured error in the motion-cued network is compared with that of proposed in

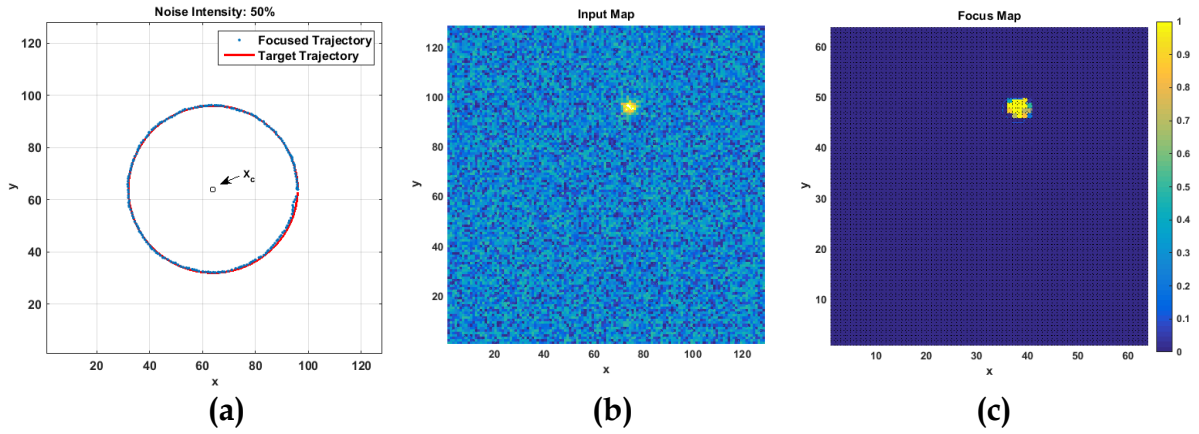


FIGURE 4-6

Noise sensitivity analysis in the attention network is illustrated. Figure (a) represents the simulated trajectory of a hypothetical input (solid red), and its respective location in the Focus-Map (blue dots). The center of the focused location within field of view is determined by weighted averaging over the activity of the Focus-Map neurons. The center of circular trajectory Xc is placed in fovea. Figure (b) shows a single frame of the Input-Map perturbated by additive Gaussian noise. The noise intensity in this example is set to 50%. Given the noisy Input-Map, the activity of the Focus-Map is exhibited in Figure (c).

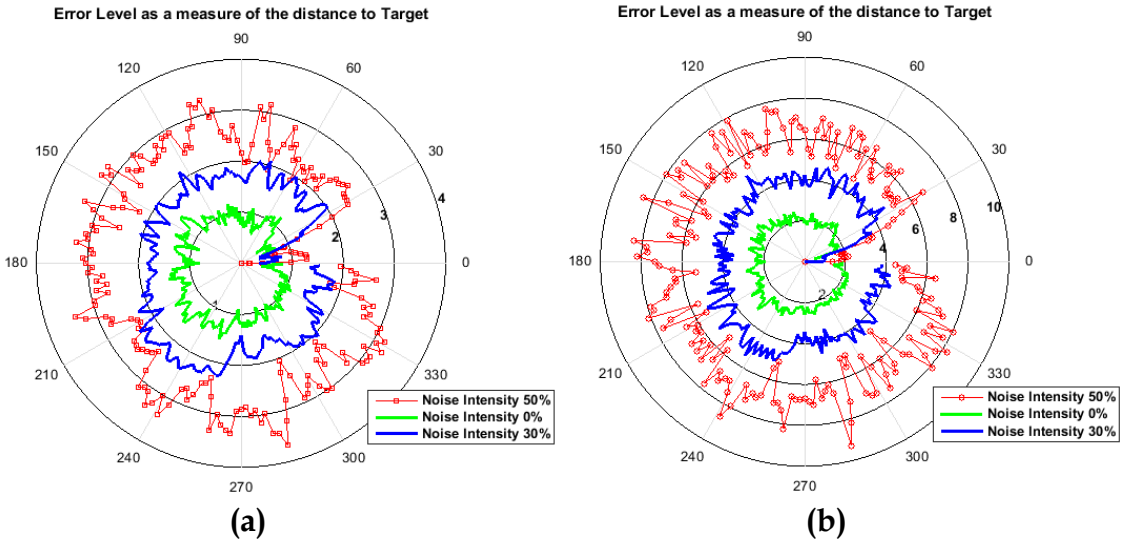


FIGURE 4-7

The distance of decoded target in Focus-Map to its original location in Input-Map is used as a measure of error. In Figure (a) and (b) the error for three different sets of stimuli are depicted within polar coordinate. At each stimulus, the level of noise is set to 0%, 30% or 50%. In (a) the performance of motion-cued network is shown, and in (b) we have shown the performance of the network proposed by [Rougier and Vitay 2006] Note that the scale of polar coordinates in (a) and (b) are significantly different.

[Rougier & Vitay 2006]. As is demonstrated in FIGURE 4-7, the integration of predictive information with direct sensory evidence in our model has significantly enhanced the robustness of the attention network against noise. On the other hand, variance of the response is also reduced as compared to the model proposed by Rougier and colleague.

4.1.3.2 Collision Scenario

Any object is in fact associated with its descriptive features (e.g. color, size) that help to recognize and thereby to point on that object. A moving object also carries motion signals including its direction of motion which help to distinguish it from distractors. The prominent role of motion cue in attention can be shown in collision scenario, where the target collides with a distractor in the scene. If the target looks similar to the colliding distractor, it is hard for the nervous system to recognize the target if it does not put the dynamics of the scene in calculation. In this experiment we have used a single moving ball as an artificial target which collides with another moving object. The size of the distractor is chosen to be twice greater than target. The moving trajectory for both objects is straight with constant velocity, starting from one corner of visual field to the other corner so that objects collide at the center of scene (near fovea). In FIGURE 4-8 the trajectory of the target and distractor are shown by red and green dashed-lines respectively. The activity of the Focus-Map neurons at each frame is also decoded to compute the focused location (blue line with square marker in FIGURE 4-8). As we can see in the Left diagram

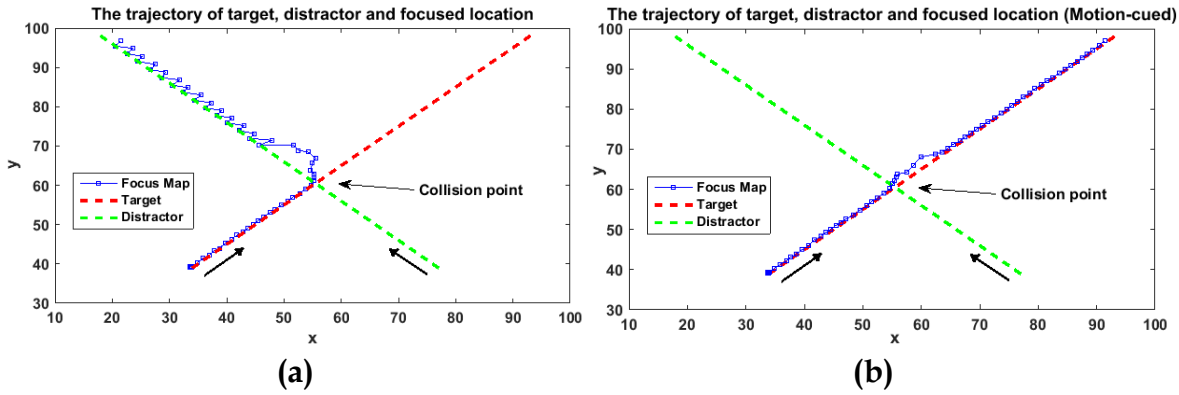


FIGURE 4-8

Performance of motion-cued attention network in collision scenario is compared with basic attention network described in [Rougier and Vitay, 2006]. A hypothetical target moves along a straight path, while it collides with a greater distractor near fovea. The trajectory of target (red dashed line), distractor (green dashed line), and the decoded location of focused object (blue line with square marker), are plotted for basic attention network (**a**) and proposed motion-cued neural model (**b**). The velocity of both objects is assumed to be 1 pixel/frame and the direction of motion is indicated by a single arrow for each object. The whole field of view is cropped into a slightly smaller and wider window for better visualization.

of FIGURE 4-8, the basic attention network is not able to keep tracking the target after collision point. That means the distractor wins the competitive process and captures the activity of the Focus-Map. One reason for this confusion is in fact the big size of distractor which takes over the excitatory area and strongly inhibits the original target. Since in the basic approach [Rougier & Vitay 2006], there is no specific descriptive information regarding the objects to be integrated within neural model, closer the distractor becomes to the target, harder that will be for network to identify the target. By adding motion-cue and fusing the predictive information regarding the location of target, the attention network can guide the attended location to the original trajectory even in presence of a twice bigger distractor. However, as is depicted in FIGURE 4-8 right, the network has been slightly distracted towards the distractor at the collision point. This problem is beyond power of the basic attention network to solve.

4.1.3.3 Realistic Data

Oculomotor-driven Visual Motion:

In Chapter 3, we show how Dynamic Vision Sensor (DVS) represents dynamic features within a stream of asynchronously generated spikes (or equivalently events). This in fact gives the sensor an advantage of exclusively extracting dynamic objects while ignoring redundant static background. For instance, those pixels that encode a moving object, or are exposed to an intensity change, will be exclusively reported (see FIGURE 3-2). Therefore, we have used this sensor to generate realistic data in order to evaluate the proposed attention network. However, we can also use data from conventional vision sensory, in cost of additional pre-processing load.

The robotic experimental setup is shown in FIGURE 4-9. It is a 6-DOF robotic head equipped with a pair of Dynamic Vision Sensor and 6 dynamixel-249 servos. The servos can precisely guide and control the rotation of DVS and the head so that the robot can simulate any arbitrary patterns of eye and head motions. Target is a laser spotlight which is blinking with frequency equal to 50 Hz. The blinking laser pointer produces a periodic intensity change at the pointed location, and thereby, generates a cluster of events every 10ms (even if it is not moving). The distractors are the magnets in different color stuck to a white background. As we can see in FIGURE 4-9, the magnets form a NST logo and it is chosen to be much bigger than target. In this experiment, rather than moving the target or distractor, we have driven the servos in such a way that the DVS moves from one corner of visual field to the other side, back and forth. This pattern of eye motion will generate an apparent visual-motion and thereby produces many distracting events (see FIGURE 4-10 left). Even though human brain is able to recognize the sensory consequences of eye motion and exclude it from external sensory information (the visual-motion that is exclusively generated by external stimuli) [Lindner et.al 2005] [Britten 2008], in this experiment we assume the entire data as a sensory-driven pattern of visual motion. In

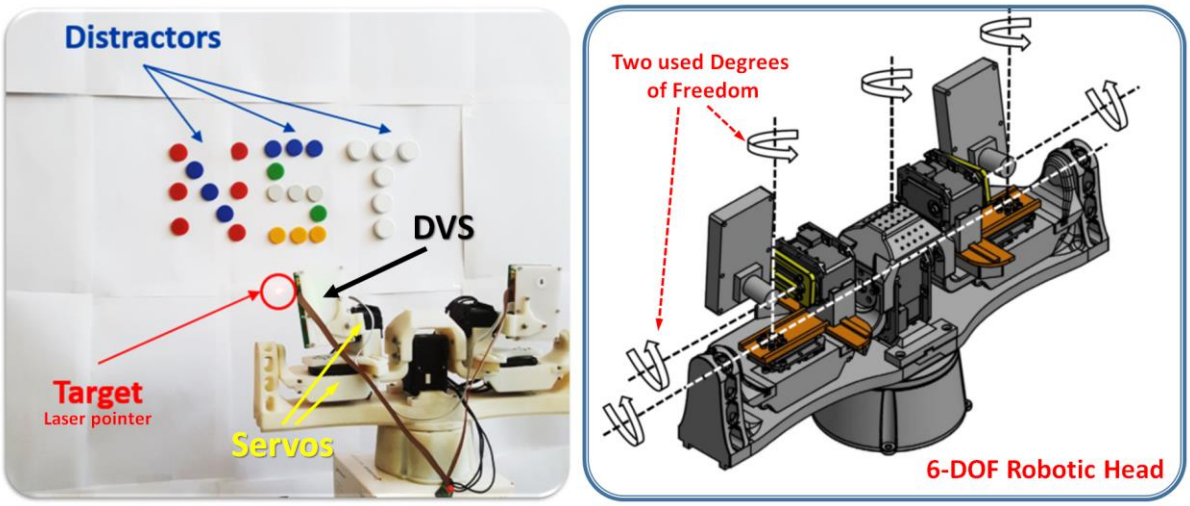


FIGURE 4-9

Left: A picture of the experimental setup is depicted in which a single blinking Laser pointer is used as a target; the colored magnets that form NST logo on a white background are used as distractor; and the DVS is mounted on a robotic-head platform. **Right:** The robotic setup is equipped with 6 precise Dynamixel servo motors that give the robot 6-DOF in order to control the rotation of vision sensors and the head. Using this configuration, it is possible to point out to any arbitrary locations in space. However, in this experiment we have just used 2-DOF of the platform to simulate the motion of a single eye. Having the static distracting objects including edges, rotation of DVS would generate a large number of distracting events.

practice this assumption is true as we exclusively would like to assess the impact of motion-cue on the attention control and not specifically to measure the artifacts generated by oculomotor action.

Left column of FIGURE 4-10 shows the activity of sensory receptors (DVS event framed for 20ms), and the middle column shows the respective activity of the Focus-Map neurons. As is demonstrated the attention network can keep an estimated location of spotlight at each frame and completely cancel out the distracting events (NST logo and edges within visual scene). In right column the predictive pattern of activity generated by motion detectors are illustrated.

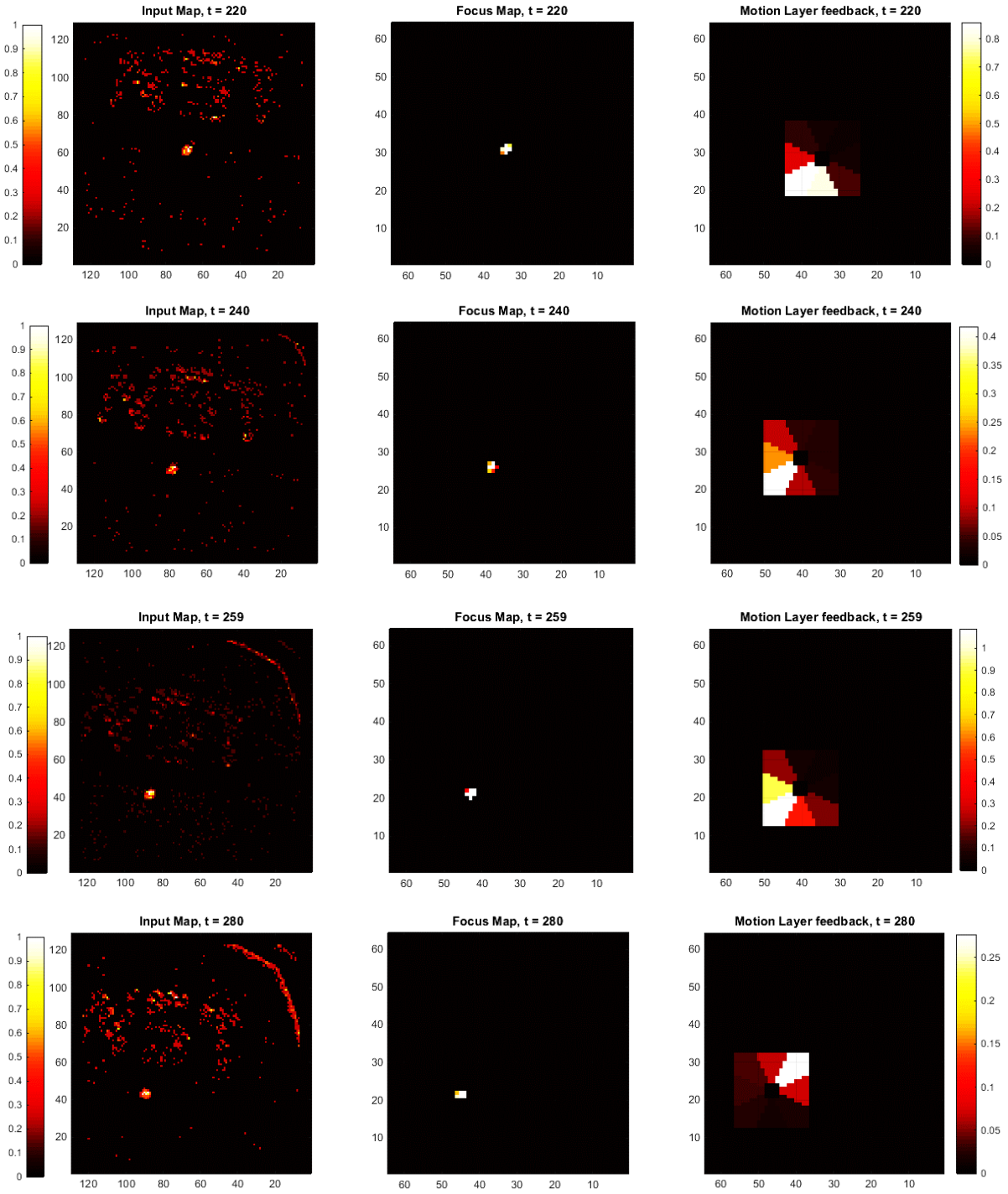


FIGURE 4-10

Left column represents the activity of Input-Map at every 20 frames starting from t = 420; **middle** is the activity of Focus-Map at each time, and **right** column shows the neural activity which is generated by motion-sensitive layer as a prediction of next location of the target, and is added to the focus-layer according to Equation (4-3). At time t = 280 (the last frame), the apparent motion of target is changed. In this experiment the sensor is moving, not the target.

Two colliding persons, a Visual Tracking problem:

To demonstrate the viability of motion-cued approach, we have analyzed the performance of the network in a realistic collision scenario. In this experiment two persons are walking through the hallway in opposite directions such that they meet at near fovea. A single static DVS records the activity of dynamic objects in the scene including artifacts. Since the data structure in DVS is asynchronous, we have framed the events every 50ms (ten times smaller than neural time constant), in order to feed input to the network. The desired outcome for this experiment is to focus on the right-side person, and to exclude the other one as a distractor. As we discuss in Section 4.2.2.2, the basic attention network fails to perform such a task (see FIGURE 4-8). Eventually, given the activity of Focus-Map at each time, we have estimated the location of attended person.

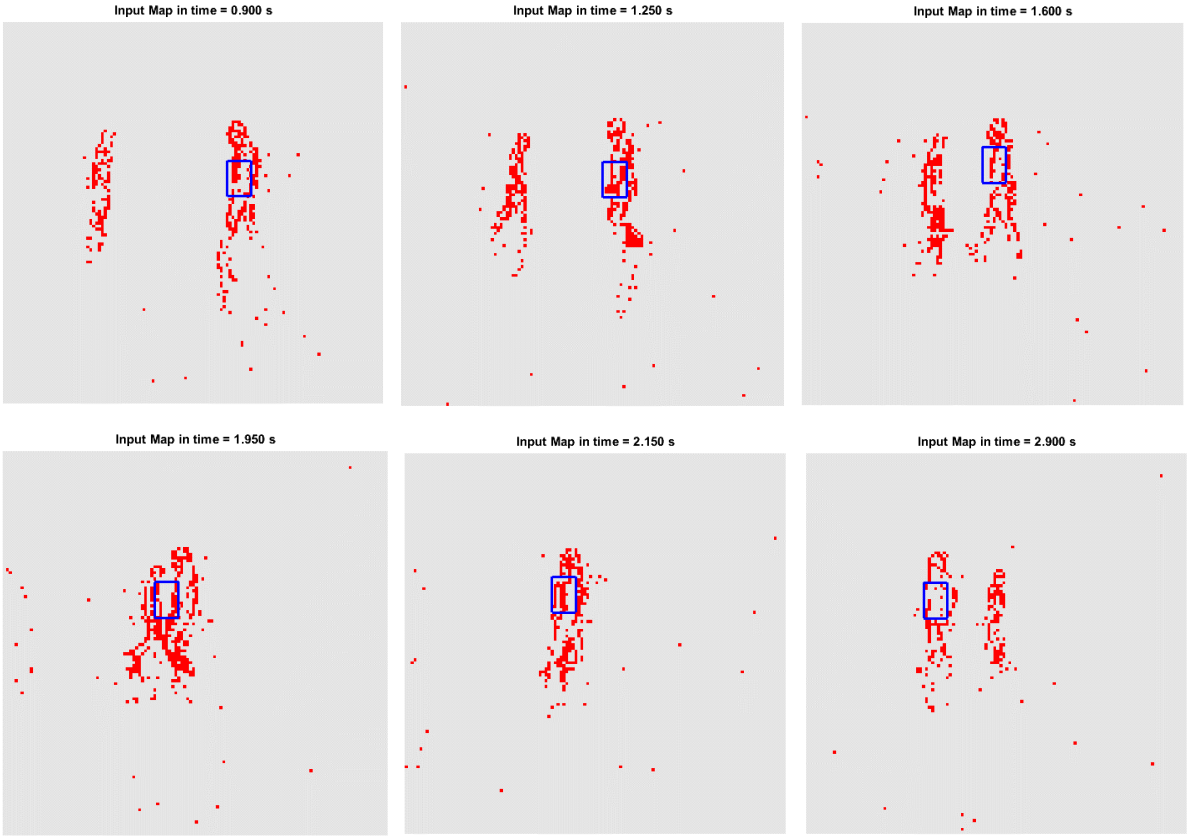


FIGURE 4-11

The performance of the motion-cued attention network is evaluated in a realistic experiment. Two persons are moving in opposite direction along a hallway and the data from DVS is recorded and framed every 50ms. Each frame provides a single Input-Map at each time step. Decoded location of the focused person (right person) is depicted by a blue rectangle.

In FIGURE 4-11, the framed DVS events or equivalently the input of the network at each time step, is shown. The decoded target at each time-step is also indicated by a blue rectangle, centered at the decoded location. As is demonstrated in FIGURE 4-11, using the proposed approach, the network is able to keep a focus on target even after collision. As a matter of fact, and cognitively speaking, we would expect the right person in FIGURE 4-11 not to change his course of walking abruptly within a fraction of second. Therefore, this gives a predictive belief about the expected location of the target at next frame. Therefore, we know that the person should show up most likely at left side of collision-site. This cognitive description is formalized within network's hierarchy and the way that integrates motion-driven predictive belief and direct sensory evidence. However, as is shown in frame t = 1.95, the decoded focused location is slightly drifted towards the distractor, as the target becomes closer to distractor.

4.1.3.4 Velocity Sensitivity Analysis in Motion Estimation Network

As we discussed in Section 4.1.1.3, the velocity of moving object in learning phase is set to 1 pixel/frame. Moreover, for the sake of simplicity, the motion detectors in our network is exclusively encode the direction and not the velocity of motion. Although it is possible to train the motion-sensitive network with different patterns of velocities, the current learning setup is sufficiently effective to modulate the activity of the Focus-Map. The resolution of a single motion detector is 45°. On the other hand, the purpose of motion estimation layer is to generate a rough prediction of target's location at next time step. The predicted location extends over a fraction of a Focus-Map neuron receptive field (see FIGURE 4-3). Consequently, and as a rule of thumb, the safety margin for estimation is $\theta \pm 22.5°$. If the error exceeds this threshold, it implies that the most active motion detector neuron is not the one that encodes the true angle of motion. A second threshold after which the true detector becomes completely silent is $\theta \pm 67.5°$. Any misclassification beyond this threshold should be strongly avoided.

In order to determine the range of velocity in which the network performs effectively, we have analyzed the error of the estimated angle in three different experimental situations. The experiments are similar to that of described in Section 4.1.3.1, but the velocity of moving ball varies from one trial to another. The reason for choosing a circular pattern of motion is in fact to uniformly evaluate the error for all possible angles of motion. The velocity is constant for a single experiment. The evaluated values for velocity are 0.5, 1, 1.5 or 2 pixels/frame. In the top row of FIGURE 4-12 the original and decoded angle of motion given by motion-sensitive layer is plotted. At the bottom diagrams the estimation error in time is depicted. Mean error for the velocity of 1 pixel/frame is 10.08° with the maximum value of 27°. This range of error puts this scenario in an entirely safe zone to generate an acceptable predictive pattern in the Focus-Map; on the other hand 95% of estimation instances are within the first threshold boundary (see FIGURE 4-13 and

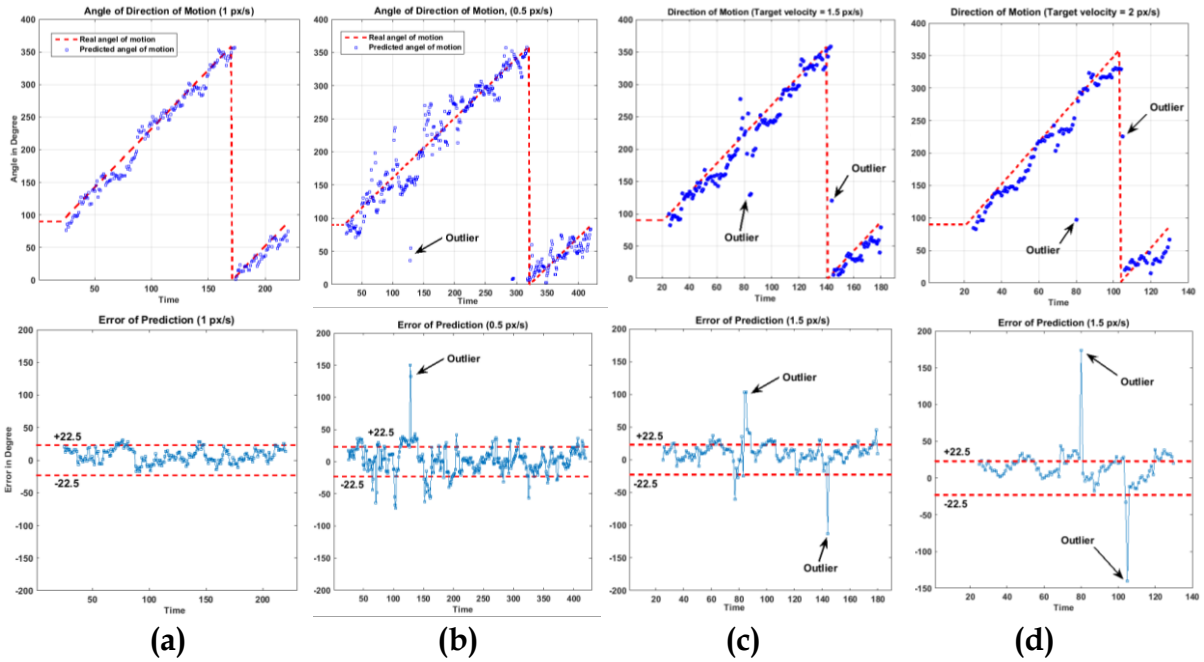


FIGURE 4-12

The performance of motion-estimation network within four different scenarios is analyzed. Similar to the experiment in Section 4.1.2.1, the object moves along a circular trajectory with a constant velocities. At each column the estimated angle of motion is compared with reference angle for the object moving with a constant velocity. Top row shows the original and the estimated angle of motion, while in bottom row the estimation error is plotted in time. The velocity of object in (a) is set to reference velocity, i.e. 1 pixel/frame which is identical to that of used in learning phase (see Section 4.1.1.3). In column (b) the velocity is half of reference velocity, i.e. 0.5 pixel/frame. In (c) velocity is set to 1.5 pixel/frame, and in (d) it is twice as large as learning velocity, i.e. 2 pixels/frame. The outliers are indicated by arrows and red dash-lines in bottom figures show the region of sensitivity for a single motion neuron, i.e. $\pm 22.5°$.

FIGURE 4-12- (a)). As is illustrated in FIGURE 4-12 (b), when the object moves twice slower, it becomes more error prone. The mean error for this experimental situation is 16.5° and the maximum error is 72°. Note that we have excluded single outliers in this case. However, more than 75% of the estimations are still ideal to use (see FIGURE 4-13). When the velocity is increased to 1.5 pixels/frame, the mean error is slightly decreased down to 16.01°, and the maximum error is also shrunk to 45°. At this experimental situation, 80% of the estimations are useful (see FIGURE 4-13). That means 0.8 of the estimated angles are within the first threshold while the rest is tolerable but still error-prone (green bar in FIGURE 4-13). Moreover, 2% of estimated angles are totally unacceptable. In the last experiment we increase the velocity up to 2 pixels/frame (as twice as reference velocity). To have an intuitive understanding of how fast that would be, the velocity of target person in FIGURE 4-13 is roughly 0.88 pixel/frame. In this situation, the mean error has

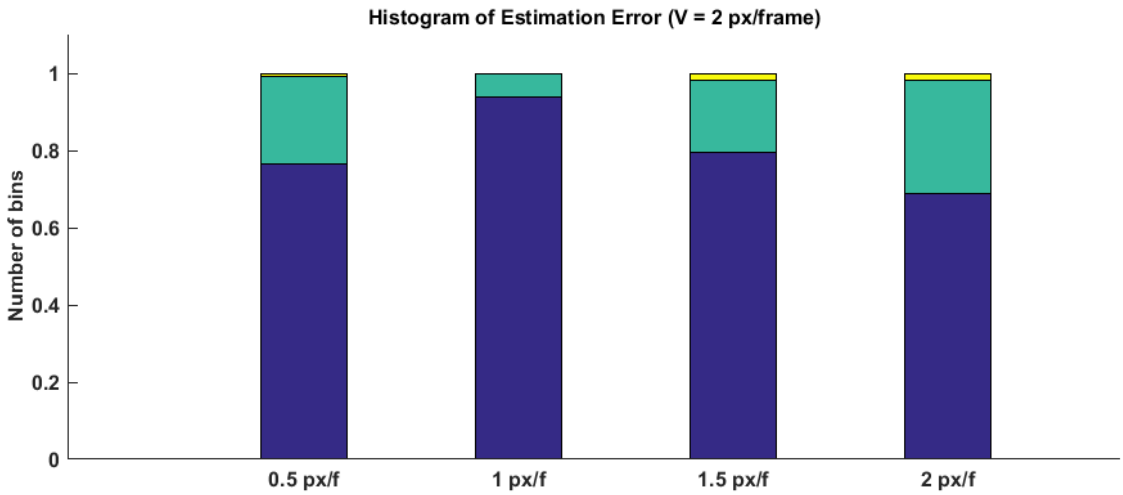


FIGURE 4-13

The stacked bar-graph shows the relative frequency of estimation instances that are sufficiently correct (blue), or exceeds the first threshold of estimation error (green), or the second threshold (yellow). The blue bar in fact indicates the percentage of desirable outcomes from motion estimation layer, within each experimental situation; green bar is the ratio of the estimates that are still tolerable even though the estimated angle is deviated from stimulated motion, but the yellow is the ratio of instances that should be completely avoided (beyond second threshold).

increased to 19.66°. The maximum error though is still comparable with other experiments and is equal to 44°. In this case 69% of estimation instances are ideal, more than 29% exceeds the first threshold but not the second one, and 2% are not useful at all. As long as these instances do not consecutively happen, that would not be destructive within the Focus-Map. If we look at the bottom diagram in FIGURE 4-13 (d), at those time bins that a significant value of error is exposed, the profile of error is impulsive and that does not form a plateau. That implies that the next estimations and thereby the corresponding generated predictive patterns of neural activity, would probably compensate the current imposed error within Focus-Map. Otherwise the attended location would be biased towards the cumulated error.

4.1.4 Remarks

One of the main functional benefits of multisensory integration is to guide behavior in case one sensory modality is not exclusively able to handle the task. In this work, we have proposed a hierarchical recurrent neural model that fuse visual motion-cue within a basic attention network. As we demonstrate in Section 4.1.3.1, adding the predictive evidence regarding the location of the target, the noise-sensitivity of the model is considerably decreased. On the other hand, situations the basic saliency-based network fails to keep

the target in a right focus field. For instance, in FIGURE 4-8 the attended location is completely deviated towards the distractor after collision. But, by adding a new evidence regarding next probable location of the target, the motion-cue helps to retrieve the focused object back to the trajectory (see also FIGURE 4-11). To train motion estimation network, the velocity of training pattern is assumed constant. In Section 4.1.3.4, we show that the estimated angle for a reasonable range of different velocities is still sufficient to modulate Focus-Map properly. It is important to note that the motion-estimation network becomes error-prone as the velocity increases compared with the trained velocity. Nevertheless, very fast motions cannot be captures by the Focus-Map either. This is because the leaky nature of neural response in this model. In better words, the evaluation of the neural dynamics in Equation (4-3) is limited to neural time constant, and time bins. For instance, if $I(x,t)$ changes rapidly in time, it cannot be represented within the Focus-Map, and thus the Focus-Map will lose the intermediate locations of the trajectory. However, it will be interesting to train motion-sensitive layer with non-constant velocity patterns; and afterwards evaluate the performance of the whole model. This can be a topic for future works. As opposed to fast motion, very slow-motion scenario is essentially not problematic even though the performance of motion estimator drops drastically in this case. Because when the target is not moving, focus field does not basically need motion-cue and input stimulus will persistently activate a single location of the Focus-Map.

To evaluate the model in real-world situations, we have performed two experiments using realistic data. We have used recorded data from DVS sensor. The reason we choose DVS is because of its superiority in capturing dynamic features of the scene and representing them in the form of quasi action potentials (for more detail check Chapter 3). The computational process of proposed model is synchronous assuming that the visual information is provided at regular time instances and evaluated periodically. Nevertheless, the photoreceptors in retina asynchronously encode the light intensity. Similarly, neurons in visual pathway are also not clocked. This principle is well accommodated in Dynamic Vision Sensors as the circuit of receptors are asynchronous and are exclusively sensitive to intensity change rather than amplitude. The synchronous style of information processing in our model is inherited from DNF[46] and imposes explicit limitations in processing time. Rougier et.al proposed an algorithm to evaluate DNF model of Equation (4-3) asynchronously [Rougier & Vitay 2011]. The problem with this method is that the behavior of asynchronous model does not necessarily resemble system dynamics in some situations [Rougier & Vitay 2011]. One possible solution to tackle this challenge is using a spike-response coding rather than firing-rate

[46] Dynamic Neural Field

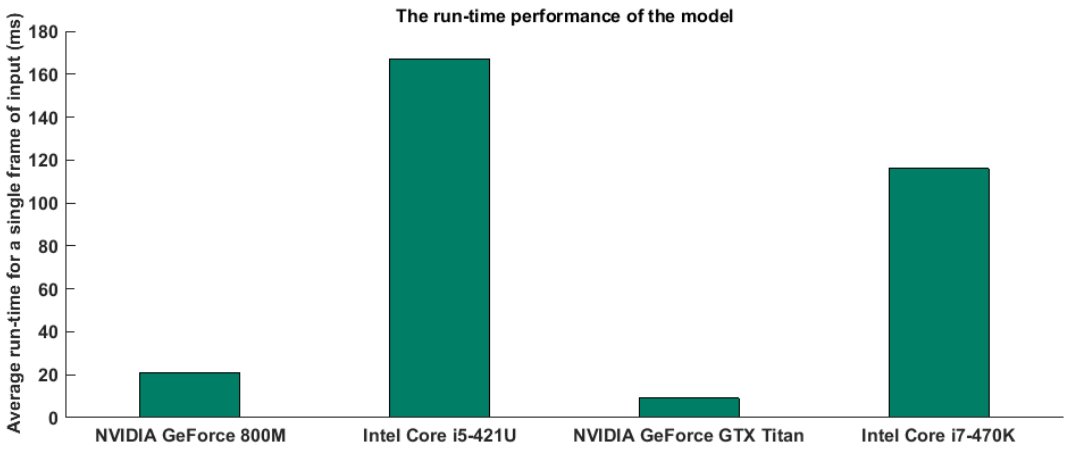


FIGURE 4-14

Average run-time of the hierarchical motion-cued attention network, within four distinct computational platforms. The processing time is measured during the experiment in Section 4.1.3.3.

coding. In [Gue et.al 2015], we have proposed a small-scale prototype of such a network in which the motion estimation layer is trained using STDP[47]. The learning scheme is unsupervised and is borrowed from [Bishler et.al 2012]. However, it is hard to scale this network up due to the limited computational resources. In [Koyuncu 2016], we have instantiated the current motion-cued model within four separate computational platforms. The run-time is measured for processing the input frames from the robotic experiment described in Section 4.1.3.3. A summary of average run-time for different platforms is shown in FIGURE 4-14. The best result is achieved by GeForce Titan with 9ms of processing time. This enables the network to evaluate more than 100 frames per second. Whereas, the stand-alone CPU, e.g. core-i7 is not clearly suitable for real-time evaluation of Equation (4-3). As a further research work, transforming this architecture into sophisticated neuromorphic embedded platforms and massively parallel machines such as SpiNNaker or TrueNorth, would be a way to tackle this impediment.

[47] Spike-Time-Dependent-Plasticity

4.2 Relation Satisfaction, Reference Alignment and Forced-Fusion using Attractor Dynamics

4.2.1. What is the Attractor Dynamics in Cortical Circuits?

Cortical mechanisms that brain employs to optimally create a coherent form of percept are context-dependent but share similar principles [Carandini & Heeger 2012]. One almost Omni-present neuro-computational principle in cortical circuit is Attractor Dynamics. A simple example of such a principle is associative memory. When we see an obscure picture of our school, after few seconds we can remember where the picture is taken. Or by looking at the mustache of a portrait we can guess who is most probably in the picture. In fact, the nervous system performs a memory retrieval process by which we can fetch the experienced moment and its associated elements from our memory, given a partial evidence of that. From another point of view, our brain creates an implicit form of our sensory experiences which guides us to restore data from noisy or partially observable sensory input. This prototypical representation of data is well formalized by theory of cortical maps [Das 2005] [Quiroga et.al 2005]. One prominent governing dynamic between cortical maps is Attractor Dynamics [Jun 1991]. Cortical Maps can be interpreted as Attractor points or prototypes in feature space and when they are activated by a noisy input, the interacting dynamics of the neurons will find an attractor point (or prototype) which corresponds best to the input signal. To put a Helmholtzian spotlight on this concept, the judgment about the location where an obscure picture is taken, might differ from one subject to another. Our personal sensory experiences exclusively shape subjective form of prototypes or equivalently attractors. For instance, Quiroga et.al studied the neural activities of the ventral pathway in few subjects, and strikingly observed that there is a "*Jenifer Aniston Cell*" for some subjects [Quiroga et.al 2005]. This cell exclusively responds to a picture of *Jenifer Aniston*. But, for those who might not know this character at all, this picture might activate parts of the memory which is associated to a friend who shares some similarities with Jenifer Aniston, e.g. hair style, eye-color.

4.2.1.1 Multisensory Convergence as a Specific form of Relation Satisfaction

In this section, we argue that coherency function in the context of multisensory inference can be formulated as a relation satisfaction problem. This function helps the perceptual system to integrate noisy signals so that they agree on a unified value according to a coherency function. We have discussed in this section that the manifold of coherency function (relation) can implicitly treated as a cortical map, and thereby the process of relation satisfaction can be implemented by Attractor Dynamics. Neurophysiological findings in multisensory neurons of Superior Colliculus support this

notion [Stein et.al. 2014]. Stein and colleagues observed that dorsal neurons of SC in cat are spatially registered within multisensory coordinates. The receptive field of these neurons are developed to align the orientation in retinal coordinate with head-centered acoustic coordinate, and body-centered somatosensory coordinate [Meredith et.al 1992] [Wallace and Stein 1997] [Xu et.al 2012]. In this sensory convergence setup, the modality-specific values of orientation are related to each other according to an identical relation function. On the other hand, the multisensory response of SCd neuron is super-additive with respect to unimodal responses [Wallace & Stein 1997]. This property is called multisensory enhancement and it is evident that cortical feedback is essential for its emergence in SCd [Alvarado et.al 2007].

There are several theoretical studies that have explained the properties of SCd neurons using Self-Organizing-Map and Neural-Field-Theory [Martin et.al 2009] [Bauer et.al 2012] [Magosso et.al 2008]. However, it is still not evident that all properties of SCd neuros are also present in cortex [Seilheimer et.al 2014]. Moreover, none of these models argued the problem of validation in Sensor Fusion (see Section1.2.2), nor incorporated the neural correlates of reliability in the model. Although, SOM-based models can well describe the development of Reference Alignment in multisensory integration, it can neither explain the stochastic dynamics of neural activities, nor the probabilistic characteristic of behavior. In this section, we will show how Attractor Network can alternatively describe the neural correlates of probabilistic world and thereby deal with validation problem. We will show.

4.2.1.2 Reference Alignment as a problem of Relation Satisfaction

Similar to multisensory convergence, reference alignment can be also interpreted as a problem of relation satisfaction, since the association between sensory cues creates a relation manifold. Sometimes the relation manifold is linear, e.g. translation of eye-centered coordinate into head-centered coordinate; and sometimes it is nonlinear, e.g. mapping of joint-angle of arms to body-centered depth cue. In his work we show how Attractor Dynamics can perform multisensory relation satisfaction and thus deal with reference alignment. We use a Hebbian Learning scheme and Divisive Normalization to train the attractor network in order to implement an arbitrary relation function between encoded sensory variables. Once the network trained, each stored pattern of neural activity within interacting populations will implement a single point of attraction. The relation manifold is in fact an attractor hyper-surface in multi-dimensional space. In addition to relation satisfaction, we have demonstrated that the network is also able to perform inference reasoning, de-noising, decision making, and reliability-based cue-integration. The latter is known as one of the main problems within Multisensory Perception – validation problem. In Section 5.3.1 we show how using Gain-File modulation as a computational principle in cortex, the dynamics of the attractor network

can be biased in favor of more reliable cue. As a consequence, the final attractor-point would be closer to the reliable cue. That means the contribution of less reliable cues can be suppressed. Along with the proposed neural framework, we have introduced a supplementary circuit that can first analyze the statistics of sensory cues, and then, preserves the relative reliability of signals. Eventually this circuit will modulate the synaptic projections of uni-sensory neurons to multisensory neurons according to the encoded reliability values. A multisensory heading-estimation experiment is performed using a mobile robot in order to validate the performance of this neural framework.

In next section we elaborate the general architecture, neural encoding mechanism, and the dynamics and the learning in the network. In Section 4.2.3 and 4.2.4 some computational abilities of the network e.g. estimation, de-nosing, cue integration and decision making are shown for a linear and a non-linear relation function. In Section 4.2.5 we demonstrate a practical heading estimation robotic application using a distributed dual-modal version of the proposed network. And finally, Section 4.2.6 summarizes and concludes this section.

4.2.2 Attractor Network for Relation Satisfaction

4.2.2.1 General Architecture and Neural Encoding

The general architecture of the attractor network for a tri-modal cue integration scenario is shown in FIGURE 4-15. The network consists of three encoded populations (R^n) and an intermediate layer (A_{lm}). Each cue is encoded by the activity of a spatially distributed population of neurons with overlapping wrap-around Gaussian tuning curves. Since intrinsic neural activity in brain is governed by Poisson variability, the initial activity or equivalently selectivity of a single neuron r_i (number of spikes per second), is drawn from a Poisson distribution with mean firing rate of neuron tuning curves, $\Phi(\kappa, x)$; see equations below where κ and σ are constant showing activity strength and width of neurons tuning curve respectively, $x^c{}_i$ is preferred value of i^{th} neuron, ν is spontaneous activity which is set to 0.1, and finally x is the input stimulus (FIGURE 4-16).

$$P(r_i|x) = \frac{[\Phi_i(\kappa,x)]^{r_i}}{(r_i)!} e^{-\Phi_i(\kappa,x)} \tag{4-8}$$

$$\Phi_i(\kappa, x) = \kappa e^{-\frac{|x-x_i^c|}{2\sigma^2}} + \nu \tag{4-9}$$

All neurons are linear threshold neurons and input neurons are reciprocally connected to intermediate layer A_{lm} ($W^n_{RA} = W^n_{AR}$). To keep input stimuli into topographically arranged spatial registers and to copy the cues into a common frame of reference, R^1 and R^2 populations (population vectors of x_1, x_2) are projected to the intermediate layer using a fixed *von-Mises* weighting distribution as following equation [Jazayeri & Movshon 2006]:

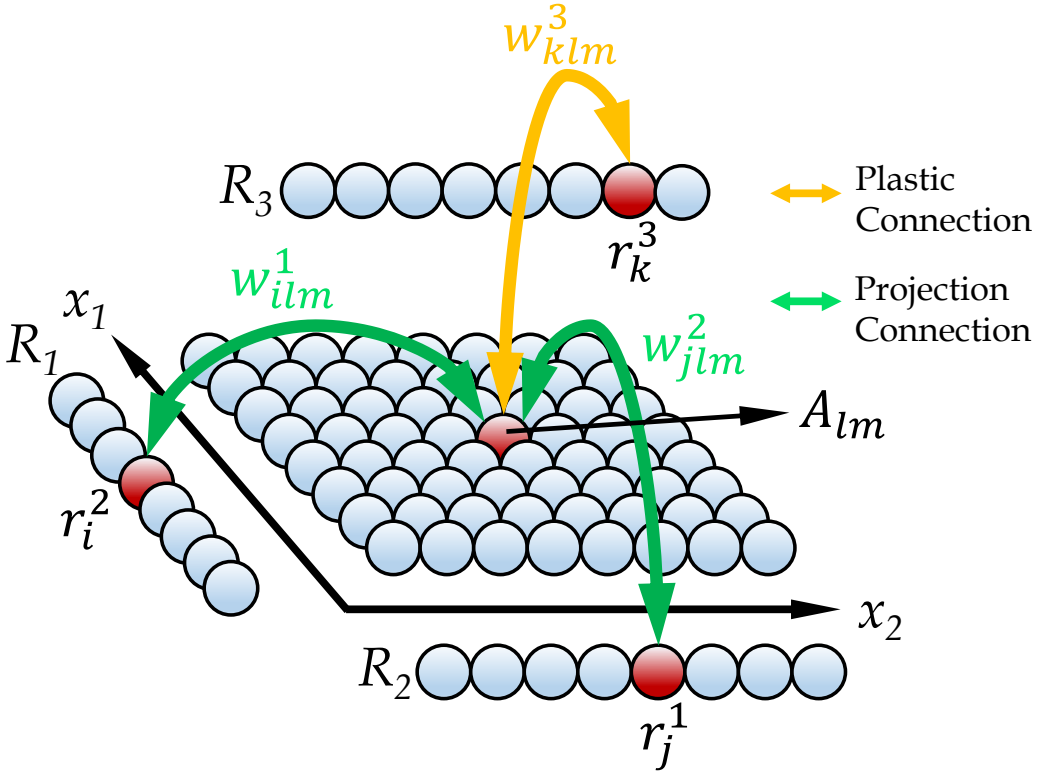


FIGURE 4-15

General Architecture of Network connectivity for three variables that are encoded using three populations. R^1 and R^2 are projected to an intermediate neural-sheet A_{lm} according to Von-Mises function. The connection of third variable x_3 to the intermediate layer is plastic so as to realize the relation function $F(x_1, x_2)$. All input populations are reciprocally connected to the intermediate layer.

$$W^1_{ilm} = e^{\frac{(i-l)\left(\cos\left[\frac{2\pi}{N}\right]\right)-1}{(\sigma_1)^2}}, W^2_{jlm} = e^{\frac{(j-m)\left(\cos\left[\frac{2\pi}{N}\right]\right)-1}{(\sigma_2)^2}} \quad (4\text{-}10)$$

Where W^n_{ilm} is the synaptic weight between i^{th} neuron of n^{th} input population (r_i^n) and lm^{th} intermediate neuron (a_{lm}), N is the number of neurons in each population and σ_n tunes width of projection. Synaptic connectivity between R^3 neurons and intermediate layer, W^3_{klm} (yellow arrow in FIGURE 4-15) is modifiable to construct the relation F by means of associative Hebbian Learning. In order to perform integration over more than three spatial cues, intermediate layer can be simply organized as a cubic or hyper-cubic topographically arranged population of neurons. Furthermore, the way of encoding and line-attraction dynamics of the network, enable us to initialize input cues, based on their relative reliabilities.

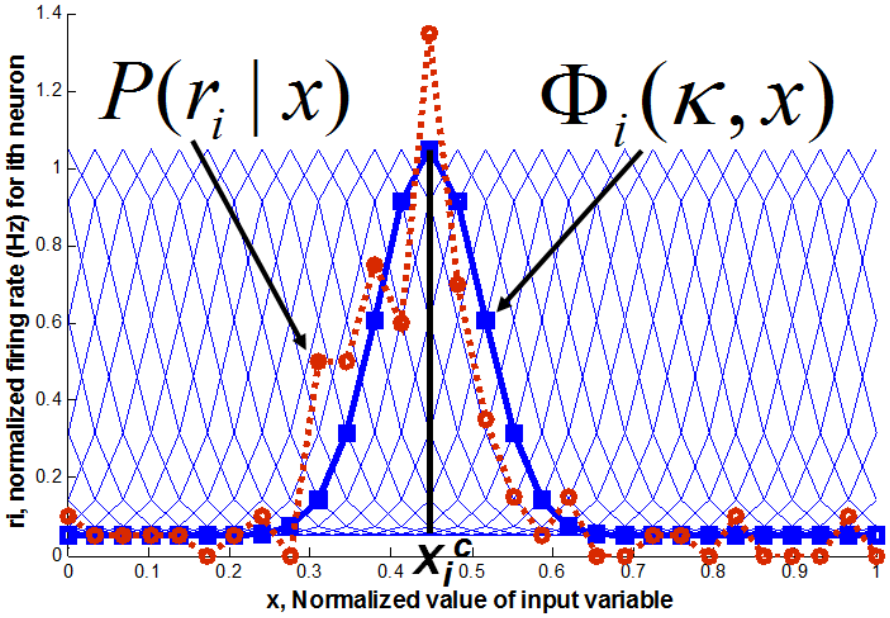


FIGURE 4-16

Red diagram shows the instant activity of the i^{th} neuron r_i (y-axis), in response to a normalized stimuli x (x axis), governed by *Poisson* variability; **Blue:** i^{th} neuron tuning curve or the expected activity (Φ_i), centered at x_i^c as the preferred value of i^{th} neuron.

4.2.2.2 Network Dynamics

Through dynamics of the network, population activities or equivalently encoded cues would be shifted so to satisfy relation function. In other word during the network's dynamics, input cues follow a trajectory to be converged toward surface of attraction in steady-state. In each time step the activity of single intermediate neuron is weighted sum of momentary activity of connected input neuron which is normalized by Divisive Normalization to keep single bumps of activities and eliminate the effect of ridge-like pattern of activities (see FIGURE 4-15). Equations (4-11) and (4-12) represent the dynamics of intermediate neurons:

$$A_{lm}(t+1) = \frac{(d_{lm}(t))^{\alpha}}{\beta + s\sum_p \sum_q \left(d_{pq}(t)\right)^{\alpha}} \tag{4-11}$$

$$d_{lm}(t) = \sum_{k=1}^{N} W^1_{klm} r^1_k(t) + \sum_{k=1}^{N} W^2_{klm} r^2_k(t) + \sum_{k=1}^{N} W^3_{klm} r^3_k(t) \tag{4-12}$$

Where a is divisive power which tunes the sharpness of normalization, β is a constant bias to prevent division by zero and $W^n{}_{klm}$ synaptic weight between k^{th} input neuron of n^{th} input population and $l_{lm}{}^{th}$ intermediate neuron. After updating the activity of intermediate layer, activity of input populations should be updated by feedback connections and DN like intermediate neurons. See equation (4-13):

$$r_i^{n\{=1,2,3\}}(t+1) = \frac{\left[\sum_l \sum_m W_{ilm}^n A_{lm}(t+1)\right]^\alpha}{\beta + s\sum_{k=1}^N \left[\sum_l \sum_m W_{klm}^n A_{lm}(t+1)\right]^\alpha} \quad (4\text{-}13)$$

It is worth to notice that for non-invertible functions, DN is not enough to elicit bumps of activity in intermediate layer, so in addition to DN an additive inhibition using a global inhibition neuron has been used to inhibit irrelevant pattern of activities in intermediate layer.

4.2.2.3 Relation Learning

As is mentioned in previous section, to construct an arbitrary relation function $F(x_1,x_2)$ between input cues, synaptic connection of third input population with intermediate layer, W^3_{klm} can be modified by a simple associative Hebbian learning. In learning phase, after projection of R^1 and R^2 into intermediate layer followed by DN and additive inhibition, a single bump of activity would emerge, and then, plastic connections would be modified as following equation (δ is learning rate):

$$W_{klm}^3(t+1) = W_{klm}^3(t) + \delta\, r_k^3\, A_{lm} \quad (4\text{-}14)$$

In each learning epoch, synaptic weights are normalized to maintain relative strength of connections and regulate overall synaptic drive received by a single neuron similar to Synaptic Scaling in biological neurons.

4.2.3 Multisensory Inference and Cue-Integration

In this section we will validate attractor network in some computational principles. The network is first trained to learn a simple linear relation function: $x_3 = x_2 + x_1$. After learning, network is initialized by noisy patterns of activity as is depicted in FIGURE 4-15 a. Also, R^1 has been initialized by two peaks of activity or equivalently two different stimuli located in different position in uni-sensory state space; one which is totally inconsistent with other cues according to relation and another is more consistent with other cues but not perfectly satisfies the relation. In the equilibrium state of the network's dynamics (after 10 epochs), activity of intermediate neurons will converge to a single bump of activity (FIGURE 4-18). This bump would generate final stabilized population vectors (FIGURE 4-17 b). As is shown in FIGURE 4-17-b the network is able to perfectly remove the internal noise. More interestingly the stimulus which is not consistent with the other stimuli has been totally removed, and the more consistent stimulus (more spatially correlated) has been strengthened (R^1 or square-red dash curve in FIGURE 4-17-a, b). The hills of activities (or equally encoded variables) are moving towards being in equilibrium point where three encoded variables perfectly satisfy the relation (FIGURE 4-17-c). In this network N is set to 40, $\beta = 0.1$, $s = 0.001$, $a = 2$ and $\sigma = 0.45$.

By initializing one of the population vectors with zero (shutting all neurons), the network can infer and retrieve the value for unknown variable that is consistent with the

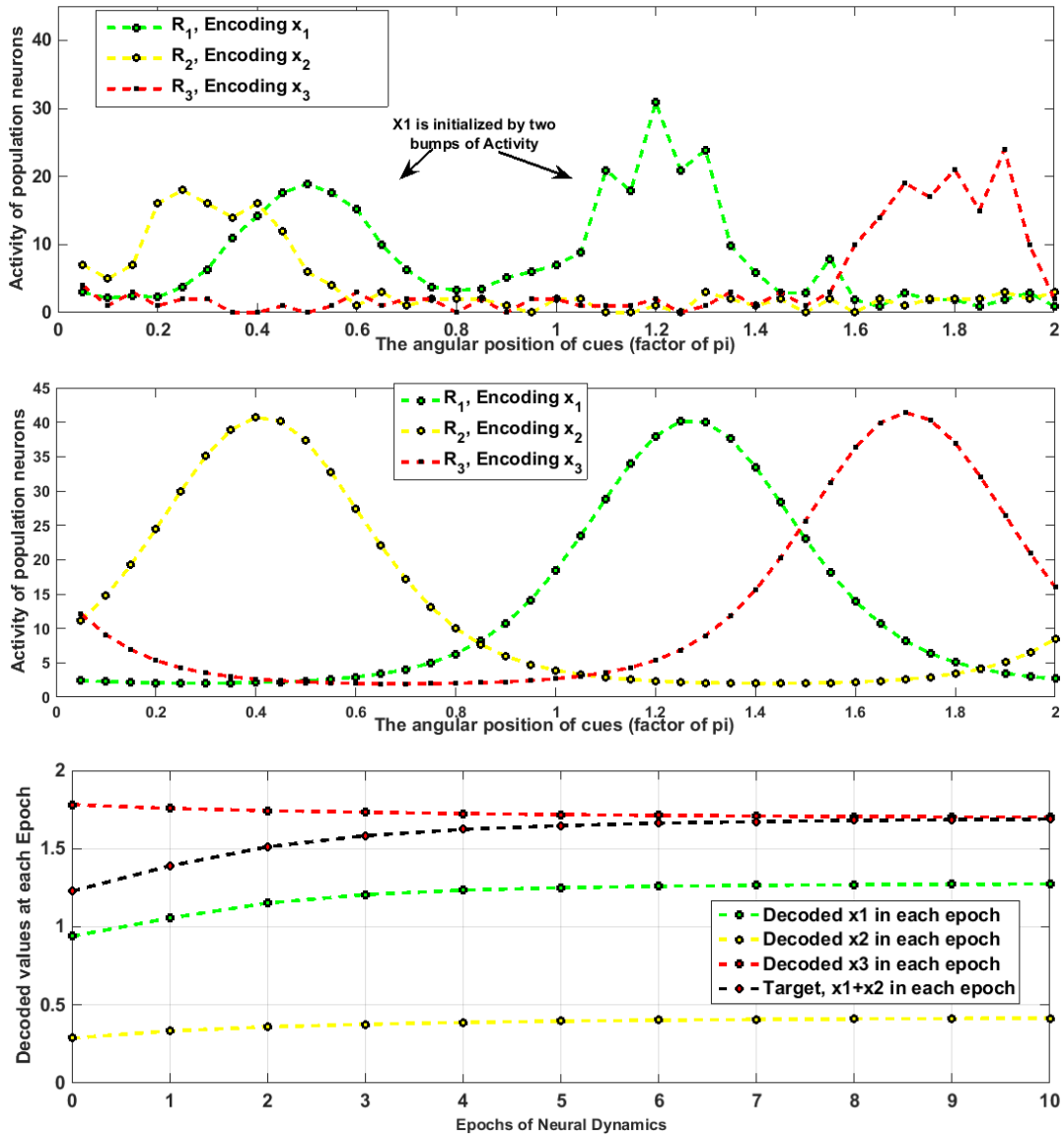


FIGURE 4-17

(Top) Initial population, (Middle) Population vectors after 10 epochs, (Bottom) Decoded values in each epoch.

other initialized variables (consistency in terms of relation). Another important feature of the network is demonstrated in FIGURE 4-16-c the less reliable cue (x_1) tends to move faster (steeper trajectory) compared to the other cues. Similarly, if one of the modalities is encoded by a smaller peak of activity (smaller κ in (2)) compared with the others, the attrac tor dynamics weights that cue as less confident cue and it would be changed faster toward being coherent with other cues with respect to relation (weighted cue integration). In section 4 by showing a realistic scenario, we will show if we perform

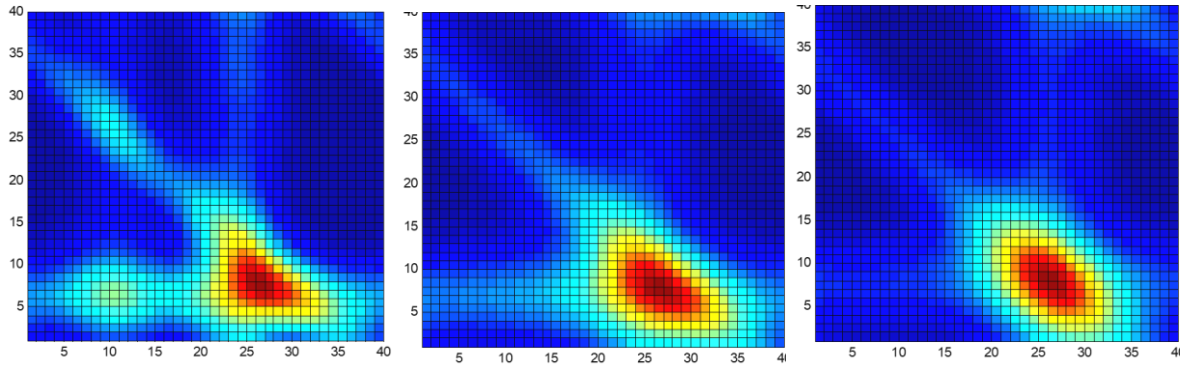


FIGURE 4-18

Momentary transient activity of intermediate neurons emerged as a single bump of activity in stable state of network dynamic, (Left) epoch=1, (Middle) epoch=5, (Right) epoch=10.

weighted encoding or equivalently weighted projection to intermediate layer, according to relative reliability of cues (e.g. reverse of Gaussian noise power in each sensory modality), the network can simply follow a near optimal cue integration.

4.2.4 Decision Making in Non-invertible Relations

In case of symmetrical or non-invertible relations like parabola function ($x_3 = x_1^2 + x_2^2$), to infer one of the x_1 or x_2 variables, it is probable to emerge two possible peaks of activity as inferred value. One solution is evaluating network dynamics and updating neuron activities using an asynchronous dynamic [8]. Another simple solution is violating the symmetry in support of one possible stimulus for unknown variable. For instance, if the network is initialized with a tiny negative bias (FIGURE 4-18 a) for the unknown cue, this

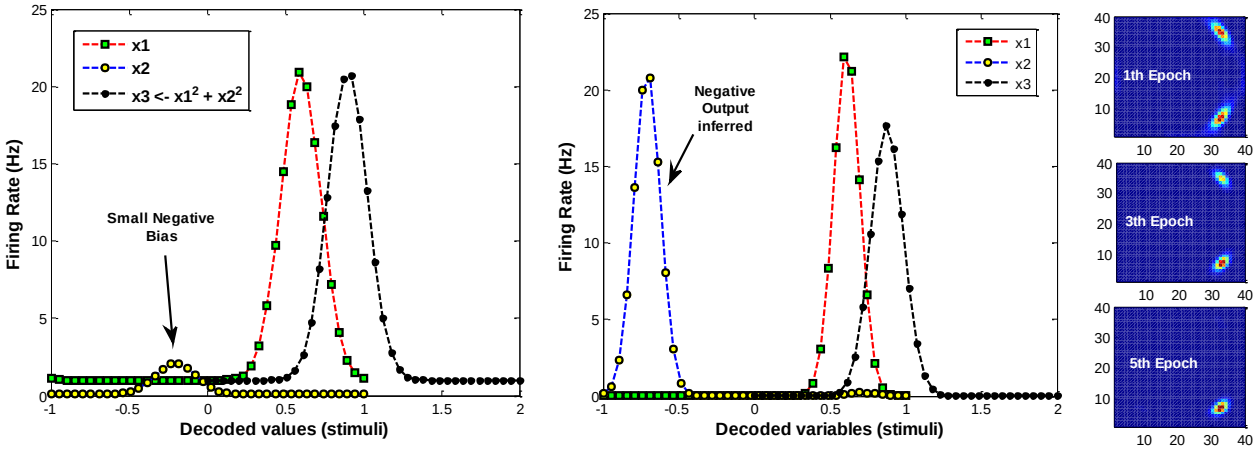


FIGURE 4-19

Relation Satisfaction for a quadratic relation function is evaluated. **Left** Initial populations, **Middle** Final populations after relaxation, **Right**: Intermediate activity in 5-epochs.

negative bias helps the network to retrieve the negative peak for hidden variable (FIGURE 4-18 b). Consequently, the bump corresponding to the positive value in the intermediate layer has been removed during network dynamics (FIGURE 4-18 c). In this network N is set to 40, β = 0.1, s = 0.002, σ = 0.38, and finally a = 3 to achieve a sharper DN inhibition for irrelevant patterns of activity.

4.2.5 Reliability-based Fusion, Heading Estimation Experiment

As a practical case study for multi-sensory cue integration, we have evaluated a tri-modal attractor network for head estimation in an Omni-direction mobile robot (see FIGURE 4-21). This network is composed of three populations R_1^g that are reciprocally connected to each other using von-Mises function of equation (4-10). A single neuron of multisensory neural ensemble R_i^g, is connected to the corresponding uni-sensory neuron in input population R_i. This synaptic connection is modulated by an inhibitory interneuron neural ensemble (shunt inhibition in FIGURE 4-20-top). This modulatory shunt inhibition is reversely proportional to the instantaneous variance of the corresponding sensory node σ_i^2. Each input population encodes a single attribute, i.e. x_i, of the stimulus according to equation (4-9). The von-Mises pattern of reciprocal connections implies that the relation function between R_i^g and R_j^g is an identical function. In FIGURE 4-20-bottom, the pattern of inhibitory synaptic connection which is generated by interneurons is shown. The mean-coding interneuron preserves the mean value of last n sensory observations. In other words, it codes the mean of sensory likelihood according to the following equation:

$$m_k = m_{k-1} + \frac{x_k - m_{k-1}}{n},\ x_k = \frac{\sum_{i=1}^{n} r_i \Phi_i(\kappa, x)}{\sum_{i=1}^{n} r_i},\ m_0 = x_0 \tag{4-15}$$

The variance in k^{th} time step is thereby coded in variance-coding neuron according to equation (4-16):

$$v_k = v_{k-1} + (x_k - m_{k-1})(x_k - m_k),\ v_0 = 0 \tag{4-16}$$

Greater value for v_k implies a less reliable sensory node, and thereby a stronger inhibitory current (denominator channel in FIGURE 4-20-bottom). The neural activity of the modulated population R_i^g is normalized using divisive normalization similar to that of formulated in (4-13).

The robotic set-up includes a mobile robot which is equipped with an IMU unit and a compass sensor. The robot explores a closed square-like trajectory in a room. The efferent copy of the motor command that drives the wheels (*odometry*) is also provided to estimate the heading angle. As a result, we have three sensory values at each time step; each provides a single estimate of the heading angle with respect to room coordinate. We have assumed that external noise is Gaussian, and the velocity of the robot is slow enough so

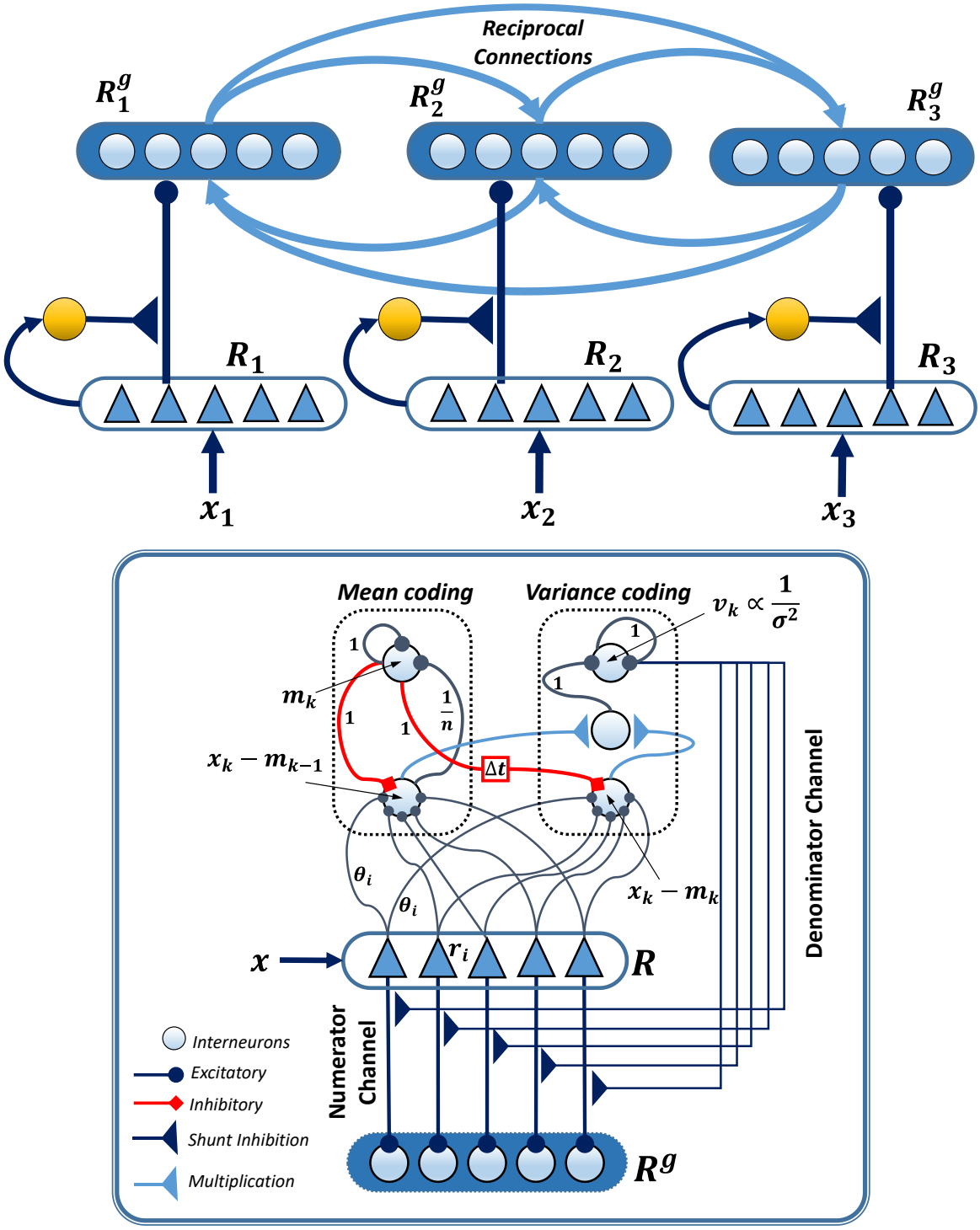


FIGURE 4-20

Top: An interacting populations are reciprocally connected using Von-Mises neural projections and the neural activity within each population is normalized using divisive normalization at each time step. Each population encodes a single sensory value. Bottom: A circuit to encode variance of sensory read-out in time and modulates synaptic projections according the relative reliability of the corresponding sensory node.

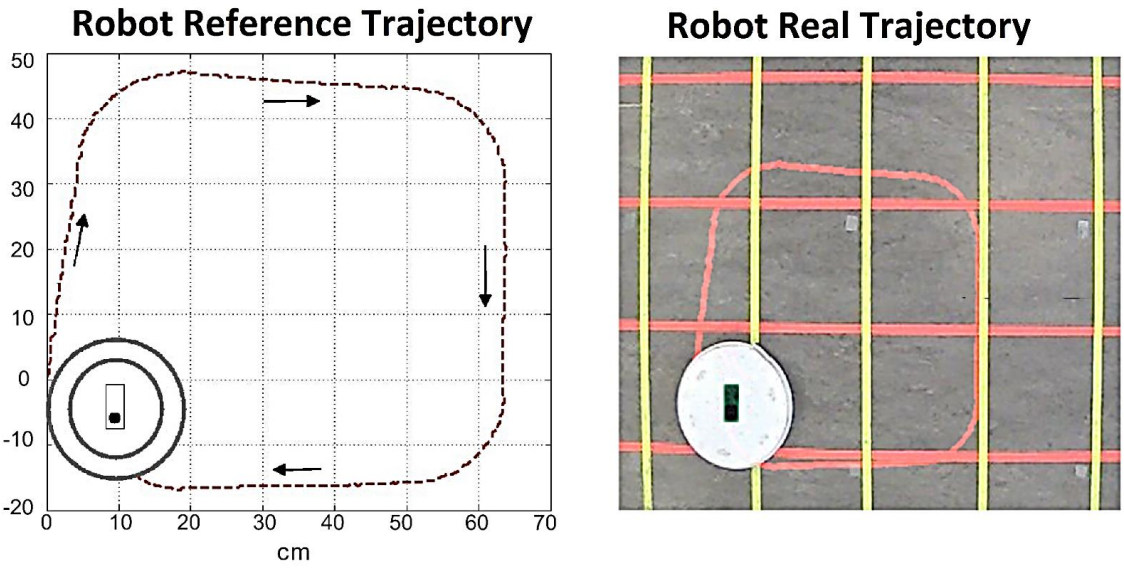


FIGURE 4-21

The reference trajectory of the omni-directional mobile Robot which explores in-door.

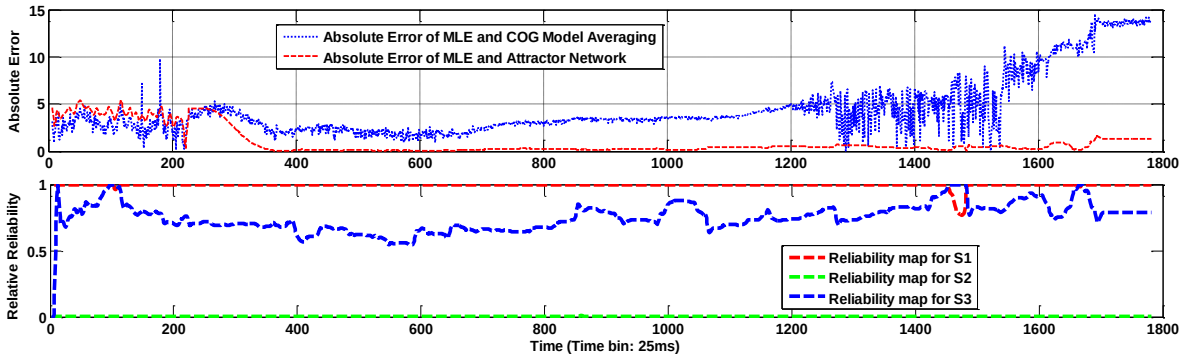


FIGURE 4-22

Top: The absolute error between MLE and COG voting integration algorithm, is compared with absolute error between MLE and Attractor Network with reliability encoding. **Bottom:** The normalized value of relative reliability of cues computed using the supplementary circuit. It is assumed that the robot moves slowly. In bottom graph S1 is Gyro (in red), S2 is Compass (in green), and S3 (in blue) is the odometery signals.

that we can compute the variance of each sensory node in time. We have computed the variance using a supplementary circuitry shown in FIGURE 4-20. This circuit implements a real-time evaluation of Expectation-Maximization algorithm for a single sensory node through a recursive process. While the robot is exploring around the space (from 0^o to 360^o) the value of sensory node fluctuation is updated and accordingly the synaptic projection of the corresponding population is modulated by a shunt inhibition. Higher the variance, stronger the inhibition. Since we want to evaluate how possibly optimality and reliability-encoding can be incorporated within Attractor Dynamics, we have

compared the network's outcome with Maximum Likelihood Estimator as a statistically optimal fusion, see Section 2.3.1. MLE combines the sensory estimates using a linear combination in which the weights are reversely related to signal variance, see equation (2-14). We also compared the outcome of attractor network with a voting-based algorithm [Triesch & Von der Malsburg 2001] [Axenie & Conradt 2013]. Simplified underlying idea of this method is that the most reliable cue is the one which is closest to Center of Gravity of all sensory estimates (for more detail check Section 2.2.3.2). In fact, the best sensory node in voting-based algorithm is the one which is more coherent with the other sensors. Two sensory nodes that exhibit similar values are in fact potentiated. Therefore, this method does not account for signal variation. In FIGURE 4-22-down the computed relative reliability for each sensory node is shown for 1780 sample points from 0^o to 360^o. For instance, it is depicted that the Compass sensor is polluted with noise more than Gyro and odometery. Therefore, the compass is assigned with a smaller reliability coefficient. This coefficient is in fact the weight of shunt inhibition in FIGURE 4-20.

In top diagram of FIGURE 4-22 absolute error between MLE as the baseline value, and vote-based algorithm is compared with Line Attractor Network (LAN). It is illustrated that the outcome of LAN network with normalized relative reliability map which is shown in FIGURE 4-22-bottom, is near optimal and close to MLE, while vote-based algorithm (COG) fails to fulfill optimality. Because it does not take into account the noise variability.

4.2.6 Remarks

The idea of retrieving information from partially reliable data using association networks is not new in machine learning. But, the dynamics of these networks is a promising and inspiring framework to understanding how cortical circuits can possibly represent, preserve and combine information to establish a coherent and robust representation of the world. We argued that each single cortical area can be interpreted as a single encoded variable, and the interaction between these areas can be interpreted as neural ensembles that are exchanging information with respect to a function. This form of interaction thereby can be seen as a special form of relation satisfaction. In fact, the neural connectivity between two cortical nodes is thought to implement a mutual relation function. This notion can simply be scaled up into the problem of reference alignment, since reference alignment is a specific form of relation satisfaction. In this section we have investigated how a simple recurrent attractor network can solve the problem of relation satisfaction and reference alignment. The network is able to learn the relation between multiple sensory cues and once it is trained, the dynamics of the network will force the sensory nodes to agree on a set of values that satisfies the relation function. On the other hand, the noise or unwanted patterns of activity (e.g. bump of activity) can be rejected as they are not in agreement with relation function. The second important feature of this

dynamics is when one of the sensory nodes is off, the other sensory nodes can predict the value of that sensory node. This resembles to the notion of Mutual Prediction (see Section 2.2.3.3). The results exhibit the capability of the network to perform de-noising, cue integration and inference even for non-invertible and smooth nonlinear functions.

We also have investigated the problem of optimality and validity in sensor fusion. Using Gain-Field Modulation, and encoding the value of reliability in neural activities, we have seen that the dynamics of the attractor can be biased in favor of more reliable cue. As a result, the contribution of the more reliable cue in the final estimate is higher. We have evaluated this approach in a tri-modal heading estimation experiment using an omnidirectional mobile robot [Firouzi et.al 2014b]. We have compared the outcome of the network with Maximum-Likelihood- Estimator (MLE) and it is shown that the network can realize a near optimal solution for reliability based multi-sensory cue integration.

Chapter 5

A Distributed Cortical Hierarchy Performs Multisensory Causal Inference, A Neuro-computational Model

"Shallow men believe in luck or in circumstance. Strong men believe in cause and effect."

— Ralph Waldo Emerson (1803-1882 AD)

5.1 Introduction

Human perception of the world strongly relies on the process of multisensory integration, as the souring world is essentially multisensory. There have been a vast amount of theoretical and psychophysical studies during last decade, regarding how the human brain combines multiple senses into a single percept [Meredith et.al 1992] [Ernst & Banks 2002] [Alias & Burr 2004a] [Alvarado et.al 2007] [Ursino et.al 2011]. In all of these works, the cross-modal singles are assumed to have originated from a single source. However, the computational strategy that human subject employs to deal with multisensory environment are not uniform as the structure of the sensory world varies in different circumstances.

Assume that we want to identify who is saying "Hi" to us, given an incoming acoustic and a visual signal (moving hands and the location of sound). Then, our motor system can program the position of our head or our eyes to say "Hi" back. One possible scenario is when both attributes are generated by a single person (FIGURE 5-1-Left). In this case combining the visual and acoustic locations into a single estimate (see Section 2.3.2) is a rational strategy that makes sense[48]. But, is this scenario always the case? If the acoustic signal comes from a source hidden from our sight (e.g., moving hands are not in our field of view), fusion of the acoustic signal with the visual location of a silent person is not rational (FIGURE 5-1-Middle). In this case, sensory cortex should employ a different and more complex strategy than fusion since the structure of the sensory world is changed.

[48] As is discussed in Chapter 1 optimality is one of the main objectives that observer wants to achieve during multisensory perception.

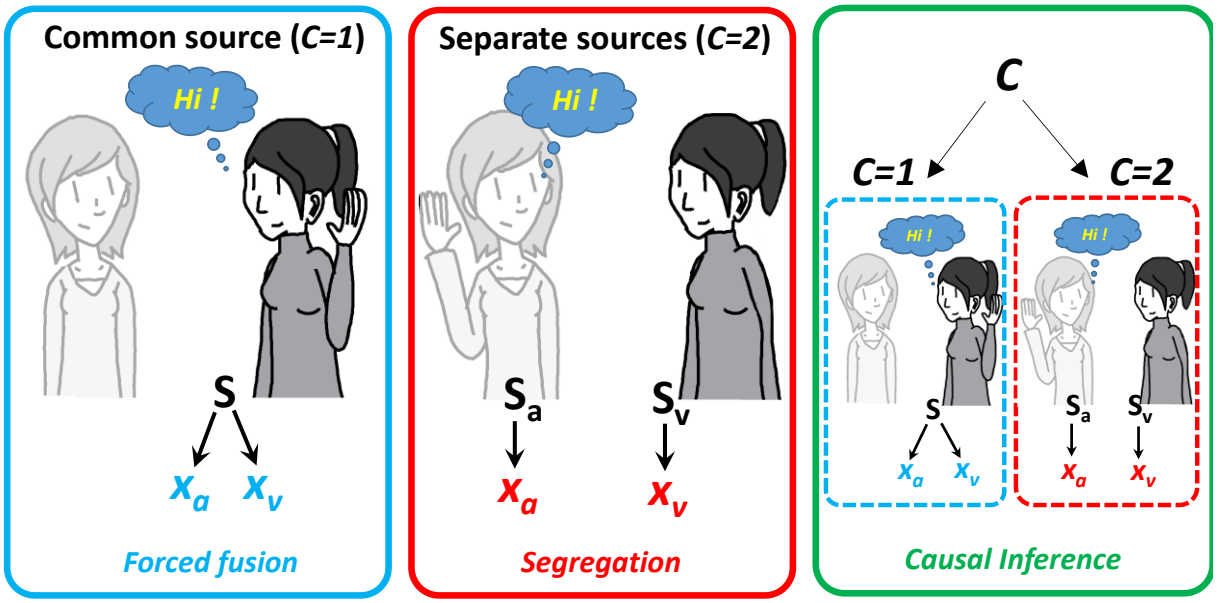


FIGURE 5-1

The process of Multisensory Causal Inference in an Audio-Visual ventriloquism paradigm is illustrated. The grayed picture represents a hidden visual stimulus that might generate an acoustic signal (saying "Hi"). So only the acoustic signal generated by the hidden person can be detected by the observer. **Left (blue box):** Given two different attributes of the world, the first existing hypothesis is demonstrated. Where the visual and acoustic attributes (x_a and x_v) both are generated by a single common source and thereby the observer should combine them into a single estimate. **Middle (red box):** The second existing hypothesis is represented where visual and acoustic signals are likely generated by two different sources and must not be fused into a single estimate. **Right (green box):** A schematic representation of the Causal Inference which takes into account the belief in an existing Causal hypothesis, in order to instantiate a proper computational strategy and eventually to estimate the location of the stimuli.

In Section 2.3.3 I have introduced a Bayesian framework for integration breakdown in which a hypothetical coupling-prior switches fusion to segregation by instantaneously widening of the width of a 2D Gaussian prior [Ernst & Di Luca 2011]. However, the plausibility of this entity in sensory cortex is questioned [Wei & Körding 2011]. Nevertheless, given the noisy sensory attributes, cortical circuit must compute the probability of existing structure in order to perform or break the fusion. For instance, in the Audio-Visual localization task depicted in FIGURE 5-1, having x_v and x_a, picked up by sensory system, the belief in whether signals are generated by a single person or by two different people, determines whether the signal must be fused into a single location, or should be segregated into two separate estimates (see FIGURE 5-1-**Right**). This process which is a hierarchical Bayesian Inference is known as *Causal Inference*. In Bayesian Causal Inference the observer infers the possible cause of the occurred multisensory event based on evidences.

5.1.1 The Problem of Perceptual Causal Inference

As we discussed in FIGURE 5-1, it is neither beneficial nor rational to integrate information from two distinct sources. Humans should always deal with multiple objects and thus multiple sources of information in the real world. Therefore, the nervous system is constantly processing the sensory stimuli across senses in an environment with a varying causal structure. The problem of applying a proper integration scenario - whether to combine signals into a single estimate or segregate them – is implicitly an inference process. In fact, we must deduce whether the current evidence across modalities correspond to the same object or not, by marginalizing the intermediate variables. This is known as the problem of Causal Inference which is not trivial to solve [Shams 2012]. Even if the signals originate from a single source, due to intrinsic noise in neural activity and environment, there is still a possibility that the observer segregates them because of a noise-driven inconsistency. To solve this difficult problem, perceptual system uses inconsistency across sensory attributes within space, time or even high-level dimensions, e.g. semantic inconsistency [Shams 2012]. On the other hand, since the human brain is an energy efficient machine (in terms of preserving information), a priori information regarding possible hypotheses should be also considered. For instance, we already know that it is not plausible to fuse the sound of barking with the image of a cat. Or the statistical frequency of sensory stimulation is usually higher in near fovea rather than in periphery [Girshick et.al. 2011].

5.1.1.1 Hierarchical Causal Inference

Along with the problem of credit-assignment and reference alignment introduced in Chapter 1, perceptual system should also solve the problem of Causality. But, the question is what components facilitate such a process in the sensory cortex? Ernst and colleagues proposed the notion of partial-fusion to model a transition from full-fusion to full-segregation (see Section 2.3.3). Despite the ability of this model to account for sensory calibration, it is doubted that the process of partial-integration functionally makes sense, because there is no clear situation at which two sensory attributes partially belong to a single source [Shams 2012]. On the other hand, the coupling-prior which reflects the joint-occurrence of sensory signals seems implausible, because it changes from trial to trial according to the momentary value of sensory likelihood [Wei & Körding 2011]. Roche et.al proposed a Bayesian framework similar to that of described in Section 2.3.2 in which the prior distribution is the superposition of a Gaussian distribution and a uniform distribution [Roch et.al 2006]. Therefore, for wide sensory conflicts the fusion will break down into segregation. This model implicitly suggests that the causal structure of the sensory space is modeled within prior. Similar to the coupling-prior model, this notion is also not plausible, and it cannot directly account for the predicted causal structure of the world. Rather, to make a direct prediction of the possible scenarios,

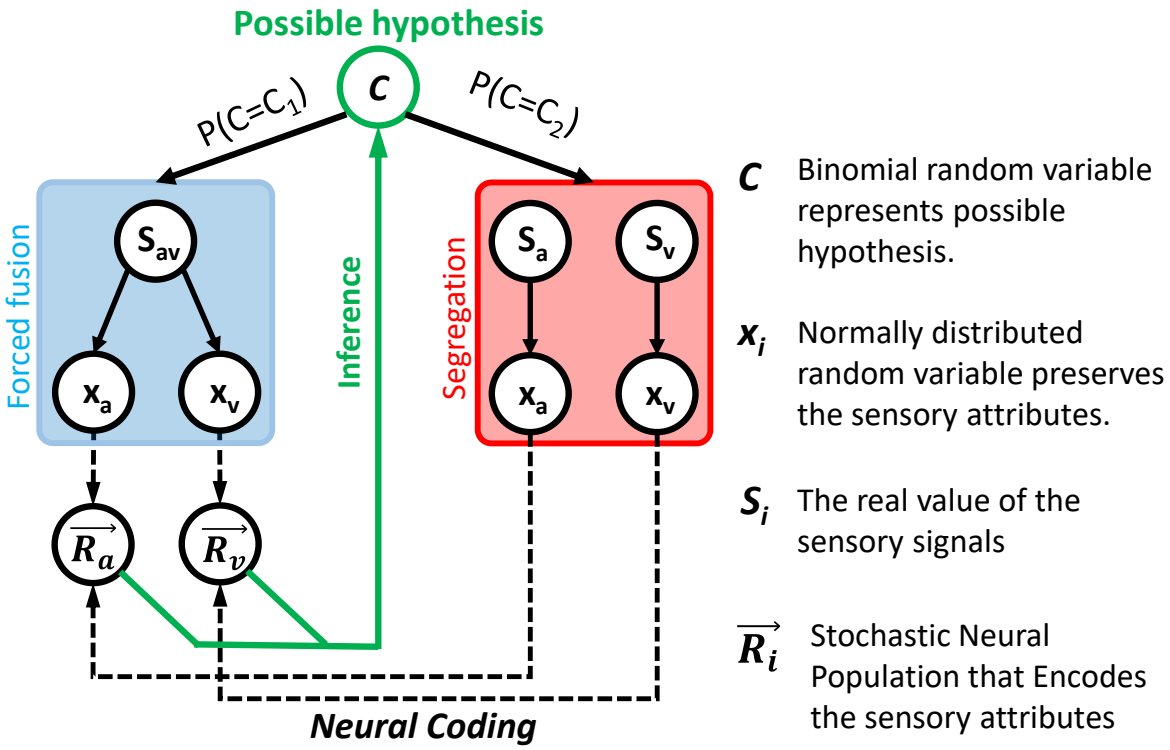


FIGURE 5-2

Generative Bayesian model of Multimodal Causal Inference in an Audio-Visual Localization task. The Bayesian Graph includes nodes which represents random variables, and arrows that reflects the conditional dependence of the random variables. It is assumed that the process of forced-fusion and segregation are handled through separate pathways as they are indicated by a blue and a red box respectively. *C* is a binomial random variable that models the probability of two possible hypotheses, i.e. common-source and separate-sources. R_i is a vector of independent random variables which reflects the stochastic neural activity in response to the noisy sensory signal x_i. The question we ask in this chapter is given the stochastic neural activities, how sensory cortex can possibly solve the problem of Perceptual Causal Inference.

Körding and colleagues suggested a hierarchical Bayesian model in which a binomial random variable explicitly describes the probability of each hypothesis and controls the process of the integration [Körding et.al 2007]. Prior to this work, Wallace and colleagues conducted an Audio-Visual localization experiment to analyze the characteristics of human responses to a cross-modal stimuli [Wallace et.al. 2004]. They noticed that the cross-modal perceptual bias often occurs in conjunction with the hypothesis perceived and reported by the subjects. In addition to that, they observed that the perceptual bias is still present even when the subjects report a common-source situation. Then, they postulate that this pattern of behavior should be generated likely within two distinct neural circuits [Wallace et.al 2004]. Following this work, Körding and colleague performed a model-based study and suggested a hierarchical Bayesian model that

includes four degrees of freedom. This model can significantly fit human data as compared to the coupling-prior model and the classical Bayesian fusion [Körding et.al 2007]. It is important to note that the remarkable fit of this model is not due to the model's degrees of freedom, because it cannot account for arbitrary data using the same degrees of freedom [Shams 2012]. A modified version of this model is illustrated in the Bayesian graph of FIGURE 5-2, wherein the random variables (signals) are shown by nodes, and the statistical dependency of the signals are shown by arrows. C is the binomial random variable which models the probability of two possible scenarios: common-source or separate-sources; S_i represents the stimulus, and x_i is the sensory evidence given the stimuli S_i. As a principle, the information that cortex preserves regarding any events in the world, is represented and transformed within stochastic neural activities, and thereby the emerged functionalities are governed by this stochasticity [Rolls and Deco, 2010]. In Section 1.1.2, we briefly discussed the benefits of this stochasticity in modeling the world. To include this principle in the generative[49] model of FIGURE 5-2**,** R_i as a modality-specific population of cortical neurons is added to the model. This neural population encodes a sensory evidence x_i. In Section 5.3.1 we describe an optimal encoding mechanism in more detail.

Having the generative model of Multisensory Causal Inference derived, the fundamental question is which cortical circuitry generates this process to compute the probability of common-cause $P(C = 1|R_v, R_a)$, and thereafter to estimate the location of the stimulus (S_v and S_a in FIGURE 5-2). To address this question, let us identify the computational components of the Bayesian Causal Inference. The first component is optimal Forced-Fusion (the blue pathway in FIGURE 5-2). In chapter 4, we discussed how attractor dynamics can execute *Forced-Fusion* in a plausible way and how a gain modulation facilitates a reliability-based cue integration. The second component is *Segregation* (the red pathway in FIGURE 5-2). This operation is done by encoding the noisy sensory evidence x_i in a probabilistic neural activity R_i. However, it is important that the encoding algorithm can well preserve the information carried by the noisy signal. In fact, one question is how a random variable can be represented through the activity of a neural population without losing information. Basically, Segregation is the consequence of separate-source hypothesis and it enforces the observer not to merge the neural activities R_a and R_v into a single estimate R_{av}. The third and the most essential element is the *Marginalization* process by which the probability of existing hypothesis should be calculated, i.e. $P(C = 1|R_v, R_a)$. In statistics, marginalization is the process of summing out the probability of a random variable, given the joint probability distribution of that

[49] A Generative Model refers to a type of learning models that describes how a pair of input-output (observed data and corresponding output) can be generated, by estimating the joint probability distribution of data. One requirement for such a model is to provide a way of sampling input-output pairs. Naïve Bayesian, and Gaussian Mixture Model are true examples of such models.

variable with nuisance variables. To marginalize out a single variable from a conditional probability, the conditional probability multiplied by the probability density function of the marginalized variable must be calculated, and the integral of the multiplication term over the marginalized variable should be computed. Beck et.al showed how possibly this process can be done in the cortex, using reference alignment and divisive normalization [Beck et.al. 2011].

Searching for a neural architecture that generates the Bayesian Causal Inference of FIGURE 5-2, we have posed three main questions:

1. First, how to represent a random variable (particularly with Gaussian-like distribution) in a pool of probabilistic neurons[50] with minimum information loss (*Encoding problem*).
2. Second, the problem of optimality in forced-fusion as one of the computational components of Bayesian Causal Inference (*Optimal Fusion*).
3. And third, how to marginalize out the intermediate random variables to compute the posterior probability of the casual hypothesis (*Marginalization problem*).

In addition to that, it is essential to determine the sensory pathways which are involved in this process. Then, we can map the components of hierarchical Causal Inference into a neural hierarchy. In the next section I will discuss a recent study that combines a model-based approach with fMRI recording, in order to map the computational components of Causal Inference to the cortical pathways, in an Audio-Visual localization task. Furthermore, we address the first posed question in Section 5.3.1, then, I discuss about the problem of optimal fusion in Section 5.3.2.1, and finally the problem of marginalization is addressed in Section 5.3.2.2.

5.2 Mapping Perceptual Causal Inference into Cortical Hierarchies, a new fMRI evidence

5.2.1. The Scope of Integration within Cortical Hierarchy

The Cerebral Cortex and particularly perceptual system is highly modular and multisensory [Shams 2012]. Surprisingly, cortical neurons that exhibit multimodal responses are found in early sensory regions e.g. V1 [Watkins et.al. 2006], A1 [Kayser et.al 2009], and S1 [Zhou & Fuster 2004]. These regions are traditionally thought to be uni-sensory. However, one must discriminate the functional role of multisensory responses in early sensory cortices which are modulatory, from that of emerged in associative areas that are supramodel. For instance, the response of a V1 neuron to an auditory stimulus is increased only when the visual stimulus is also present and the auditory signal is spatially

[50] The neurons are assumed to be Poisson Neurons (see Section 5.2.1).

and temporally congruent with it [Kayser et.al 2009]. Whereas, multimodal neurons in the Intraparietal Cortex can respond to visual motion independent of inputs from other modalities [Makin et.al 2007]. In fact, the role of IPS neurons is integration in functional level in order to execute a specific process, i.e. combining motion estimations across senses and creating the peripersonal space. This functional difference highlights the different scope of the integration through cortical circuits. However, this functional hierarchy does not imply that the sensory signals must be first pre-processed independently in early regions, and later the well-digested data is sent up to the convergence zone, as is suggested in [Calvert et.al 2004]. This primitive notion is strongly doubted today, as recent studies remark that thalamic afferents might bypass early sensory cortices in certain circumstances and through the white matter in such a way that the stimulus can be processed directly in multisensory regions [Linag et.al 2013]. In addition to that, electrophysiological recordings exhibit cross-modal neural responses in brainstem just after 15-30ms of stimulus onset [Musacchia et.al. 2006]. These results suggest either a parallel processing of multisensory signals along cortical hierarchy or underscore the important role of feedback connections from poly-sensory to uni-sensory regions [Paraskevopoulos and Herholz 2013] [Mesulam 1998]. In other words, the necessary steps to process multisensory signals does not strictly follow the functional hierarchy of cortical circuits as there are likely parallel pathways from thalamus to cortical poly-sensory regions (a direct thalamocortical channel of information, and an indirect channel through primary cortices). This parallel style of information transmission clearly leads to a rapid reaction time especially for salient events (see Chapter 2). Imagine that the acoustic and visual signals must travel across 10-12 layers of neurons to reach posterior parietal cortex and intraparietal sulcus. That requires at least 100-120ms to create a unified estimate of the stimulus location, in addition to extra tens of milliseconds to program motor output. This time scale is not compatible with recording data, even if we consider the interplay between short-term memory and perception [Linag et.al 2013] [Wozny et.al 2008] [Shams et.al 2005].

Having the governing principles of multisensory perceptual Causal Inference determined, in this chapter I have proposed a distributed neural model that can replicate this process and reproduce human data. This work is in fact the first plausible neural model on its own kind that can describe the process of multisensory Causal Inference. The main motivation of this work is to put a mechanistic spotlight on understanding the underlying neural mechanism of multisensory integration in the brain, and in general the theory of decentralized multisensory integration in sensory cortex [Yua et.al 2015] [Zhang et.al 2016] [Olcese et.al 2018]. In the next section the detailed architecture of this model is described. I will also address three main questions that I posed in last paragraph of Section 5.1.1.1, regarding neuro-computational mechanism of perceptual Causal

Inference. In Section 5.4 experimental results are described, and finally some remarks and conclusions are discussed in Section 5.5.

5.2.2. Mapping cortical regions into computational components

A new study reveals that the components of the Audio-Visual Causal Inference (introduced in Section 5.1.1) are preserved within a distributed hierarchy in cortex [Kayser & Shams 2015]. Moreover, it is argued that these components, i.e. forced-fusion pathway as well as marginalization and segregation pathway, are computed in a parallel way [Rohe & Noppeney 2015]. To identify specific regions of cortex that perform each elements of perceptual casual inference, Rohe & Noppeney developed a new hybrid approach that utilizes fMRI data and the generative model of FIGURE 5-2 [Rohe &

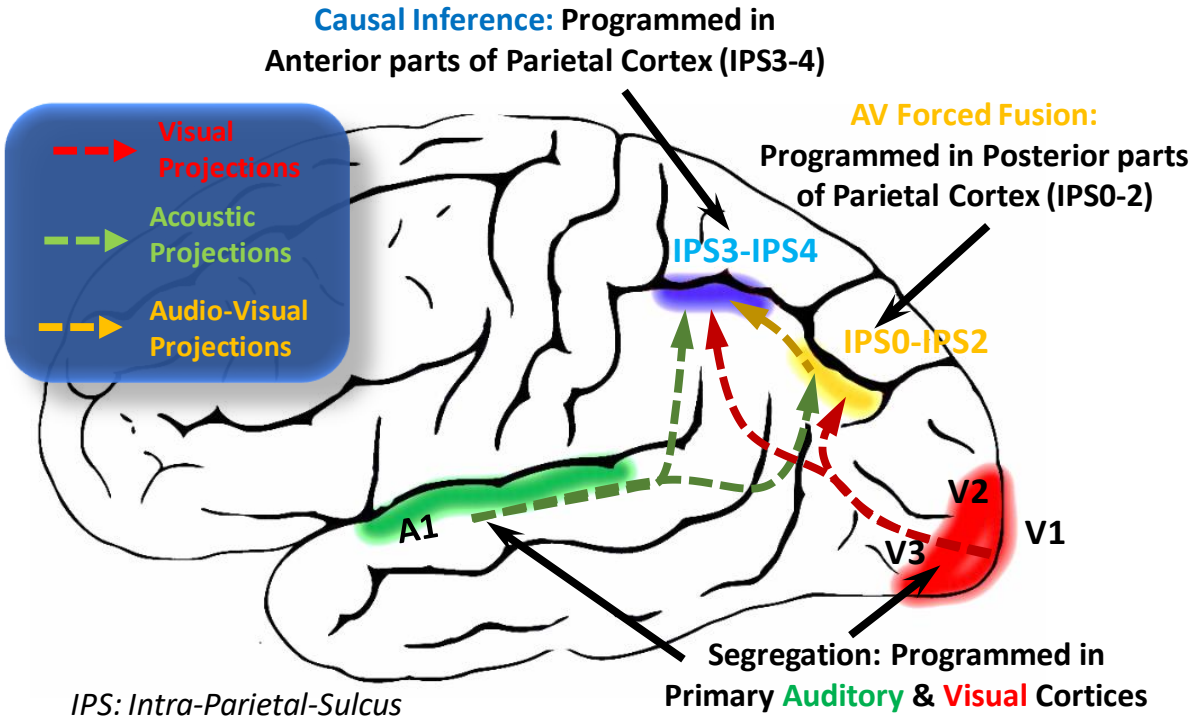


FIGURE 5-3

The identified regions along cortical hierarchy that execute the computational components of perceptual Causal Inference in Audio-Visual localization task [Rohe and Noppeney 2015]. The segregation process is represented within modality-specific primary sensory cortices (striate cortex is colored by red, and early and secondary auditory cortices are colored by green). Orange-colored region (IPS0-IPS2) illustrates the area which performs forced-fusion. Model-averaging estimate (the estimate takes into account the causal probability), is performed by Anterior parts of IPS (blue colored area). This estimate is the weighted average of full-fusion and full-segregation estimates based on perceived probability of causal hypothesis. The arrows show the flow of data and thus the neural projections within cortical hierarchy.

Noppeney 2015]. In this study, six subjects were asked to report the location of the visual stimuli (a cloud of white dots on a black screen), or the location of a brief burst sounds, and the causal situation. BOLD[51] responses from fMRI images are also collected at each session. Then, the voxel responses are decoded using a Multi-Variate-Pattern-Analysis technique [Haxby et.al 2014], in order to specify the cortical region that is most likely activated by a specific computational component. During each session of the experiment, all cortical areas along visual and auditory pathways are activated. Therefore, to determine the most-probable area that computes the perceived location, a Bayesian model-selection technique is used. First, the Bayesian generative model of FIGURE 5-2 is fitted to participants responses. And thereafter, given the spatial estimate of this model, the belief in which cortical voxels represent that estimate is calculated. The region that exhibits higher probability than other areas (exceedance probability) is chosen as the identified cortical area that instantiates the estimated location (for more details see [Rohe & Noppeney 2015]).

FIGURE 5-3 shows a schematic view of the results achieved in this work. It is depicted in this picture that a distributed circuit generates the process of multisensory Causal Inference in audio-visual localization task. As is expected, when the signals are spatially incongruent, the perceived positions of stimuli are programmed in primary sensory areas. This is because these areas predominantly reflect the uni-sensory signals and as is discussed in Section 5.2.1, they are only a little effected by other modalities. On the other hand, regions that are involved in creating cross-modal spatial maps, i.e. posterior intraparietal sulcus, computes the force-fusion estimate (orange-colored region in FIGURE 5-3). Regardless of how likely signals correspond to a common-source, IPS0-IPS2 merge the spatial evidence fetched by different sensory systems, according to the relative reliabilities of the signals (see Section 2.3.1 and 5.1.2.1). Finally, anterior parts of IPS, i.e. IPS3-IPS4 encode the spatial estimate which is provided by Causal Inference Model. In contrast to full-fusion which is a linear model, Causal model is a nonlinear combination of sensory likelihood, and the belief in causal origin of sensory observations.

5.3. Method

5.3.1. Encoding Signal Variability in a Population of Poisson Neurons

A perceptual system must represent and internalize the physical properties of an environment. Physical variables are subject to intrinsic variability, and thus, what the nervous system should deal with is a pool of probabilistic information that should be

[51] Blood-Oxygen-Level-Dependent activity is a metric in fMRI recording that reflects the response of neural ensembles in fMRI images.

characterized within neural circuits. On the other hand, a random variable is not always directly observable and there is only an evidence associated with it. For instance, given the stimulus S_{av} in FIGURE 5-2, x_a and x_v are two pieces of evidence regarding the location of stimulus that are observable by sensory system, and thus are converted into neural activities R_a and R_v. In other words, what sensory system provides for perceptual system (including cortical and subcortical regions), regarding the location of S_{av}, is described by R_a and R_v. The process which formulates this conversion is known as the *problem of encoding* in sensory perception [Simoncelli 2009]. Now, the question is how to convert an analog random variable x_i to R_i in a plausible way? Which constraints and requirements must be fulfilled during this process? Shadlen and colleagues trained two monkeys performing two-alternative-forced-choice decision task with a set of 10 visual cues (shapes). Each cue is associated with a specific reward in favor of one alternative choice [Yang and Shadlen 2007]. Strikingly, they found that the instantaneous firing rate of some neurons in LIP is directly correlated with the value of log-posterior-odds assigned to the present visual cue. It is not clear whether LIP converts information about shapes into probabilistic values or presumably neurons in ventral pathway provide that as an input to LIP? Nevertheless, this study exposes for the first time the capability of brain to extract probabilistic quantities from symbolic stimuli. How cortical neurons can learn and represent uncertainties in general is still unclear [Pitkow et.al 2015] [Stuphorn 2016]. However, within the last decade there has been a large body of research regarding how intrinsic stochasticity of neural activity in sensory cortex can possibly reflect the variability in sensory signals [Jazayeri & Movshon 2006] [Ma et.al 2006] [Simoncell 2009] [Fischer 2010] [Wei & Stocker 2012] [Ganguli & Simoncelli 2014]. In addition to that, there are relatively few theoretical studies that highlighted the underlying synaptic mechanisms account for probabilistic reasoning in cortex [Soltani & Wang 2010]. In the problem of encoding, optimality is a key requirement to obtain. That means the information content of the converting signal must not be lost during encoding and decoding[52]. Recently, it is theoretically shown that the shape of tuning curves in a pool of Poisson-like neurons can facilitate an implicit way of representing prior and sensory likelihood within a unified neural framework [Ganguli and Simoncelli 2014]. In general, there are two prominent common assumptions in all of these coding algorithms:

1. First, the variability of neural firing rate is governed by a Poisson process which seems plausible, since it can reasonably describe the stochasticity of action potentials in cortical neurons.
2. Second, it is assumed that the firing rate of one neuron is statistically independent from the activity of other neurons. Jazayeri and Movshon argued that the noise correlation is not fixed and varies from stimulus to stimulus in cortical circuits and hence it is not reasonable to deal with the input correlation structure [Jazayeri &

[52] In many circumstances, Information maximization is equal to Mean-Square Error minimization.

Movshon 2006]. Then, they concluded that under this assumption the decoder can be still near-optimal.

To describe the problem of optimal encoding in the context of sensory perception, let's assume that we have an arbitrary random variable s that is encoded by a population of n Poisson neurons $R = \{r_i | i = 1, .. n\}$ (see FIGURE 5-4). Each neuron has a specific tuning curve that is the mean activity of that neuron in response to a single variable $f_i(s)$. Give assumptions I and II, the posterior probability of the random variable s is as follows:

$$\mathrm{P(s|R)} \propto \mathrm{P(R|s)} = \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(r_i|s)}\,; < \mathrm{r_i} > = f_i(s) \tag{5-1}$$

$$\mathrm{P(r_i|s)} = \frac{\mathrm{e}^{-\lambda}}{\mathrm{r_i!}}\lambda^{\mathrm{r_i}} \Rightarrow P(R|s) = \prod \frac{e^{-\lambda_i}}{r_i!}{\lambda_i}^{r_i} = \left(\prod \frac{1}{r_i!}\right) e^{\sum \lambda_i} e^{\sum r_i \log(\lambda_i)} = \phi(R) e^{\overrightarrow{F(s)}.\vec{R}} \tag{5-2}$$

Where λ_i is the mean activity of r_i and thus is equal to $f_i(s)$, and $F(S) \in R^n$; $F_i = \log(f_i(s))$. Since in a homogeneous neural coding area under the curve of $f_i(s)$ is not varying amongst neurons, so $e^{\sum \lambda_i}$ is constant and independent of s [Ma et.al 2006], and consequently $\phi(R)$ is also independent of encoded variable. Moreover, if we assume a Gaussian-like tuning curve for neurons, log of $f_i(s)$ becomes quadratic. As a result, the posterior probability of s will be Gaussian in which the exponent component is a linear proportional to $\mathrm{r_i}$. This linear combination leads to an interesting feature: if the activity of neural population is amplified by a factor of g, equivalently the variance of the Gaussian process shrinks by a factor of $1/g$. This relation enables the coding algorithm to implicitly incorporate signal's fluctuation within the amplitude of neural activities [Ma et.al 2006]. Note that we have not imposed any quantitive assumptions regarding the tuning width or the preferred values of the neurons so far. Assuming a Gaussian function for the tuning curve of ith neuron $f_i(s)$, with an amplitude equal to gA (g is the gain-modulation factor), tuning width of σ_{tc}^i (with a preferred value equal to s_{tc}^i), equation (5-2) can be reformulated as follows:

$$f_i(s) = \mathrm{gA}e^{-\frac{\left(s-s_{tc}^i\right)^2}{2{\sigma_{tc}^i}^2}} \Rightarrow \mathrm{P}(R|\mathrm{s}) = \phi(R) e^{\sum r_i \left[\log(gA) - \frac{\left(s-s_{tc}^i\right)^2}{2{\sigma_{tc}^i}^2}\right]} = \Psi(R) e^{\sum -r_i \frac{\left(s-s_{tc}^i\right)^2}{2{\sigma_{tc}^i}^2}} \tag{5-3}$$

The expectation value of $\mathrm{P}(R|\mathrm{s})$ is equal to the root of $\frac{\partial P(R|s)}{\partial s}$ that can be formulated by the linear decoding of equation (5-4). The variance of $\mathrm{P}(R|\mathrm{s})$ is also formulated by (5-5).

$$\frac{\partial P(R|s)}{\partial s} = 0 \Rightarrow \mu_{P(R|s)} = \tilde{s} = \frac{\sum \frac{s_{tc}^i}{\sigma_{tc}^2} r_i}{\sum \frac{1}{\sigma_{tc}^2} r_i} \tag{5-4}$$

$$\frac{1}{{\sigma_{P(R|s)}}^2} = \sum \frac{r_i}{\sigma_{tc}^2} \rightarrow {\sigma_{P(R|s)}}^2 = \frac{1}{\sum \frac{1}{\sigma_{tc}^2} r_i} \tag{5-5}$$

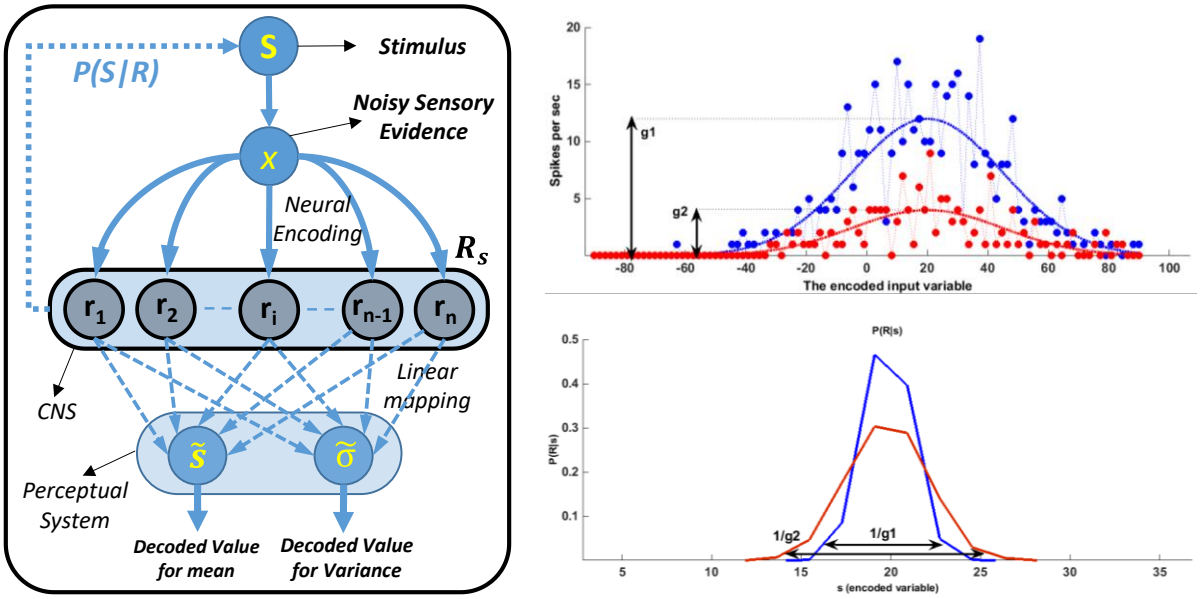


FIGURE 5-4

Left: the neural encoding mechanism of Probabilistic Population Code, for a given stimulus *S* and the noisy observation of that *x*. This encoding scheme is able to preserve the statistical properties of the random variables, i.e. $\tilde{S}, \tilde{\sigma}$, that can be directly decoded by perceptual system using a linear mapping. *R* is a vector of *n* Poisson neurons with Homogeneous Gaussian tuning curves, and arbitrary tuning width and preferred values. **Right:** Given the random variable $x = S + \eta$ *(η is an additive Gaussian noise),* on top figure the instantaneous neural activity of two encoding populations are shown. Each population is modulated with a different values of gain: g_1, and g_2. To analytically show the optimality of PPC, we perform a Monte-Carlo simulation with 2000 encoding-decoding samples. At each sample, first *x* is drawn from a normal distribution and encoded using two PPC. Each PPC is composed of 180 Poisson neurons with arbitrary tuning width and preferred values S_{tc} regularly arranged from -90 to +90. Then the decoded values $\tilde{S}$ are binned and normalized in Right-figure below. As is demonstrate in this diagram the normalized histograms are Gaussian-like whose width (reflect the variance of $\tilde{S}$) are reversely proportional to g_1 and g_2. On the other hand, the mean of both profiles are equal to 20. This shows the capability of PPC to incorporate the variance along with expectation value of a Gaussian random variable in an optimal way [Ma et.al 2006].

Since $\tilde{s} = \mu_{P(R|s)}$ in (5-3) is the mean of posterior distribution (or likelihood), this decoding is an optimal estimate of encoded variable. To analytically evaluate the characteristics of the Probabilistic Population Code, we have performed a Monte-Carlo simulation with 2000 samples. The frequency of the decoded values is plotted in FIGURE 5-4 – **Right**. As is demonstrated, the variance can be simply incorporated within neural population as a gain of neural activity. This is because of the characteristic of Poisson-neurons in which a higher amplitude implies a higher signal to noise ratio. As we discussed in chapter 4, gain modulation is one of the canonical forms of highlighting information in human brain, sometimes as a result of top-down attentional modulation

[Bisley 2012], and in some cases to reflect the coherency or quality of the sensory signals [Fetsch et.al 2012]. In our neural model, we have encoded the noisy signals according to equation (5-2) with Gaussian tuning curves.

5.3.2. Neural Model Architecture

As is we discuss in Section 5.1.2.2, the proposed neural model for the process of audio-visual causal inference is structured within a distributed hierarchical circuit. FIGURE 5-5 shows the general architecture of this model. As is color-coded in this picture, the circuit is composed of three distinct pathways: full-fusion (blue pathway), full-segregation (indicated by red), and the process of marginalization and causal inference which is colored by green. The input of the model - including fusion and marginalization

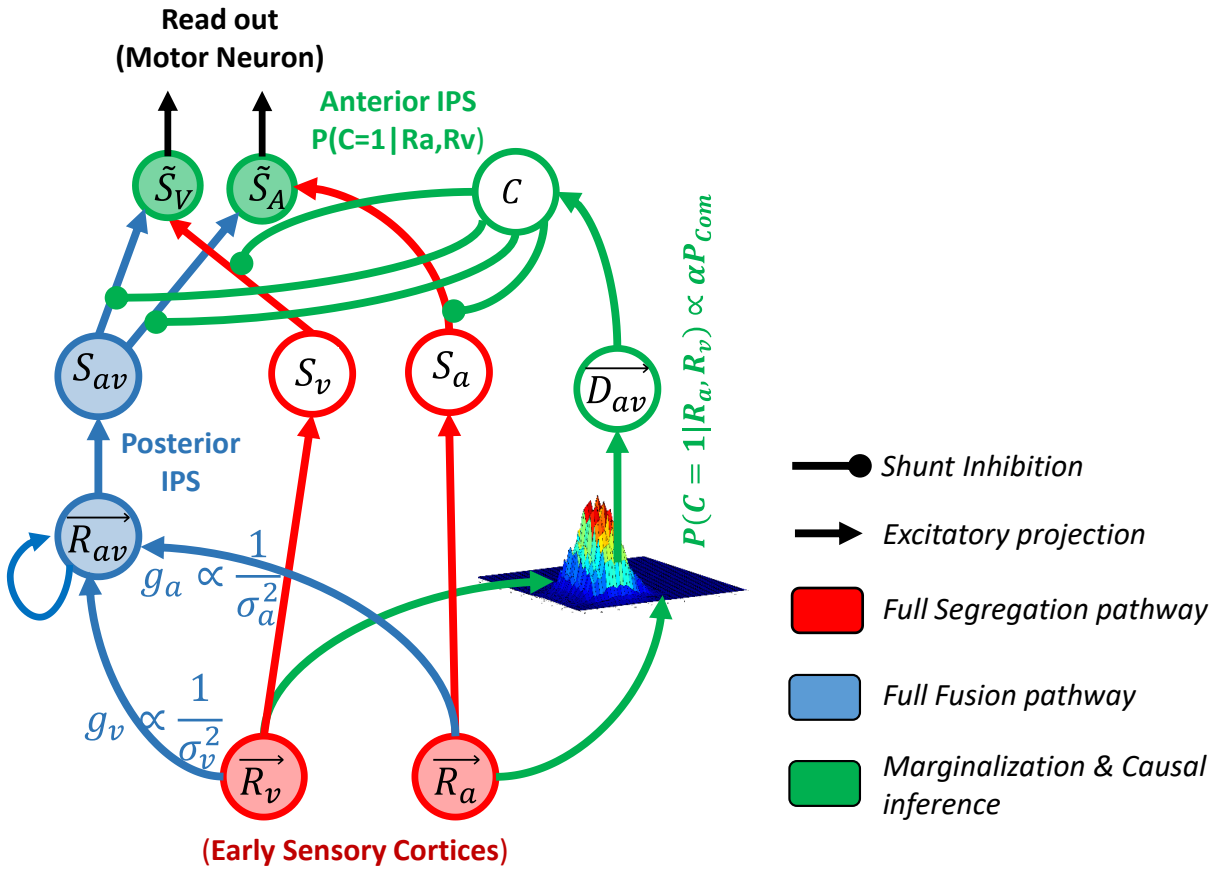


FIGURE 5-5

The general architecture of distributed neural model for audio-visual causal inference. The circuit is composed of three pathways each conducts one of specific components of causal inference: blue indicates forced-fusion model which extends over posterior parts of IPS Segregation (which is preserved within early sensory regions), is shown by red neural ensembles. And finally the process of causal inference is colored by green [Firouzi et.al 2016].

pathways - is the information encoded in early sensory cortices, i.e. R_v and R_a in FIGURE 5-4. This is the encoded sensory evidence which is fetched by sensory system and must be processed by perceptual system. The computational outcome of common-source scenario, i.e. fusion pathway is represented within R_{av}. This circuit is very similar to that of discussed in Section 4.2.3. The nervous system has no direct access to the sensory signal and thus needs to marginalize out the intermediate variables. That results in computing the belief in the present causal hypothesis. This process is handled within marginalization pathway where the probability of common-source is coded in neuron *C*. the activity of this neuron controls and mediates the flow of information from early regions to higher order areas. The synaptic projections from early sensory areas (segregation estimates) and posterior-IPS (fusion estimate) to Anterior-IPS (causal estimate) are modulated by the activity of neuron *C* using shunt inhibition.

The perceptual system relies on inconsistency across sensory attributes in order to compute the belief in current scenario [Shams 2012]. Inconsistency facilitates a mechanism to marginalize nuisance variables. However, the statistical parameters of the sensory signals e.g. covariance, prior, etc. must be incorporated in the model, whether implicitly or explicitly. Synonymously, in order to program neuron C in our model, the spatial disparity as an intermediate cue is computed and is represented in D_{av} neural population. Thereafter, neuron C maps the perceived disparity into a probability value. It is important to note that, one requirement for such circuitry is preserving the information content of the signals. Since disparity is the superposition of two random variables x_a and x_v (each encoded by R_a and R_v), the variance of disparity signal decoded from D_{av}, i.e. σ_d^2, must be approximately equal to $\sigma_a^2 + \sigma_v^2$. In Section 5.2.2.2, we will discuss how spatial disparity is optimally computed and thereby the belief in causal hypothesis is inferred within causal decision pathway. Furthermore, we will see how forced-fusion pathway optimally combines noisy attributes into a single estimate.

5.3.2.1. Forced-Fusion Pathway

Forced-fusion circuit combines multisensory attributes into a single estimate. One requirement is that the final estimate must be optimal. In chapter 2 we describe a liner fusion model which is optimal if it complies with three assumptions: noise process in each modality must be independent, additive, and Gaussian-like. The generative model of FIGURE 5-2 remarkably fits human data under these assumptions [Körding et.al. 2007]. Accordingly, in our model, it is assumed that sensory noise is governed by an independent additive Gaussian process. The forced-fusion circuit is shown in FIGURE 5-6 (a) where uni-sensory neurons r_a^i and r_v^i are projected to r_{av}^i neuron. Neurons with identical index e.g. i are tuned to identical preferred value S_{av}^i. This pattern of synaptic connection emphasizes the principle of Hebbian associativity in which the synaptic

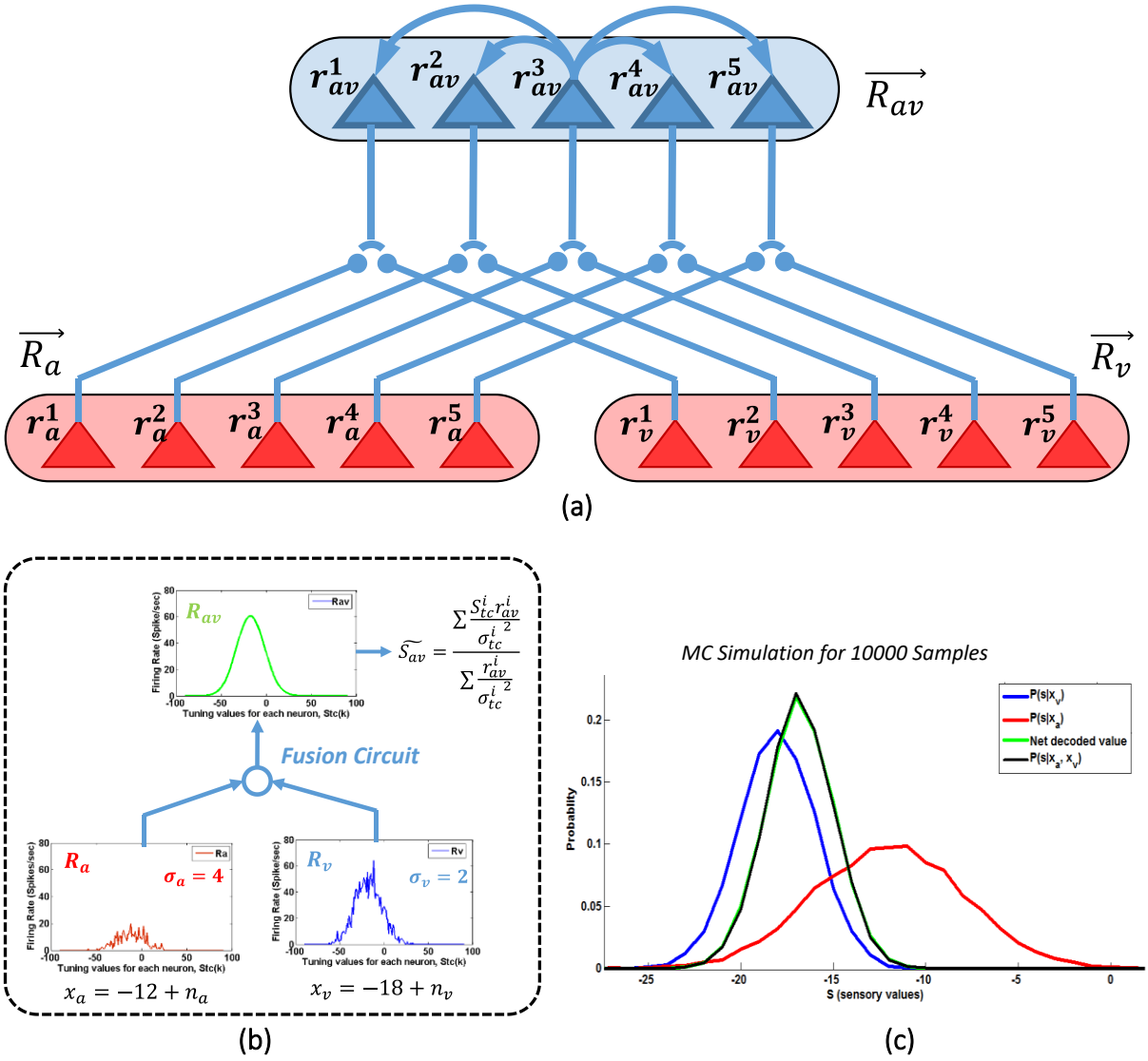


FIGURE 5-6

(a) The structure of the full-fusion pathway is illustrated in which uni-sensory neurons r_a^i and r_v^i are projected to the correlated multisensory neuron r_{av}^i. In fact the neurons with identical tuning values are connected. The neurons in R_{av} are also connected using a Mexican-hat function. **(b)** A schematic representation of a single trial of Monte-Carlo simulation. **(c)** The normalized probability distribution functions of uni-sensory evidences, fusion circuit estimates, and a Naïve Bayesian estimator.

strength of neurons with correlated activities should be potentiated. As a result, those neurons that share similar preferred values exhibit stronger synaptic strengths. Another reason behind this synaptic projection is the issue of optimality in integration. Let's assume $r_{av}^i = r_a^i + r_v^i$; the superposition of two Poisson variables r_a^i and r_v^i, is another Poisson random variable with $\lambda_i^{av} = f_i^a(s) + f_i^v(s)$. Substituting λ_i^{av} and $r_{av}^i = r_a^i + r_v^i$ in (5-3) lead to following equation:

$$\mathrm{P}(R_{av}|\mathrm{s}) = \prod \mathrm{P}\left(r_a^i + r_v^i|\mathrm{s}\right) = \left(\prod \frac{1}{(r_a^i + r_v^i)!}\right) e^{\Sigma(f_i^a + f_i^v)} e^{\Sigma(r_a^i + r_v^i)\log(f_i^a + f_i^v)} \quad (5\text{-}6)$$

1. By expanding the right-side of (5-6):

$$\mathrm{P}(R_{av}|\mathrm{s}) = \binom{r_a^i + r_a^i}{r_a^i} \left[\prod \frac{1}{(r_a^i)!} e^{\Sigma f_i^a} e^{\Sigma r_a^i \log(f_i^a + f_i^v)}\right] \left[\prod \frac{1}{(r_v^i)!} e^{\Sigma f_i^v} e^{\Sigma r_v^i \log(f_i^a + f_i^v)}\right] \quad (5\text{-}7)$$

2. Since $\sum f_i^v$ and $\sum f_i^a$ are constant, and if we assume similar tuning properties for correlated unimodal neurons $f_i^a + f_i^v = f_i^a \left(1 + \frac{g_v}{g_a}\right)$, equation (5-7) can be re-written as follows:

$$\mathrm{P}(R_{av}|\mathrm{s}) = \mathrm{P}(R_a + R_v|\mathrm{s}) = K_a K_v \, P(R_a|s) \, P(R_v|s) \quad (5\text{-}8)$$

   This proves the optimality of this approach. Note that likelihood is not a normalized probability function.
3. Moreover, as is analytically proved in Section 5.3.1 the variance of the encoded variables are reversely proportional to gain factor and $g_{av} = g_a + g_v$, which leads to $\frac{1}{\sigma_{av}^2} = \frac{1}{\sigma_a^2} + \frac{1}{\sigma_v^2}$. This equation is another good signature of optimal fusion. In chapter 4 we analytically show that even using a fully-connected network, the fusion pathway can be remarkably near-optimal. The neurons in R_{av} are laterally connected using a Mexican-hat function. Neural activity of R_{av} neurons are also normalized according to (5-9) in which N is the number of neurons, and u_k^{av} is the normalized output of kth neuron:

$$u_k^{av} = \frac{r_k^{av}}{1 + \frac{1}{N}\Sigma r_k^{av}} \quad (5\text{-}9)$$

To analytically evaluate the optimality of forced-fusion network, we have performed a Monte-Carlo simulation. At each time, a pair of random normally-distributed variables $x_a \sim N[\mu_a = -12°, \sigma_a = 4]$ and $x_v \sim N[\mu_v = -18°, \sigma_v = 2]$ are drawn and encoded using equations (5-2) and (5-3). The gain factor of each neural population is chosen to be reversely proportional to the variance of the respective sensory attribute. Thereafter, the encoding PPCs are combined into R_{av} using full-fusion circuit (FIGURE 5-5 (b)). Finally, the decoded values of final estimates are calculated using equation (5-4) and are binned within a histogram (FIGURE 5-5 (c)). In FIGURE 5-5 (c), the normalized frequency (or equivalently the probability distribution) of the final estimates of full-fusion circuit (light-green), is compared with the outcome of an optimal Bayesian estimator (dark-green). It is demonstrated that the forced-fusion circuit is perfectly optimal [Firouzi et.al 2016].

5.3.2.2. Marginalization Pathway

In Bayesian generative model of FIGURE 5-2 if we assume that sensory signals are perfectly noiseless, i.e. $S_v = x_v$ and $S_a = x_a$, a deterministic discrimination threshold e.g. D_{th}, can help the perceptual system to judge whether the signals originate from a single source or not. For example, if $d_{th} = 10°$ and $|x_a - x_v| = 15°$, so $d_{av} > d_{th}$, then the subject

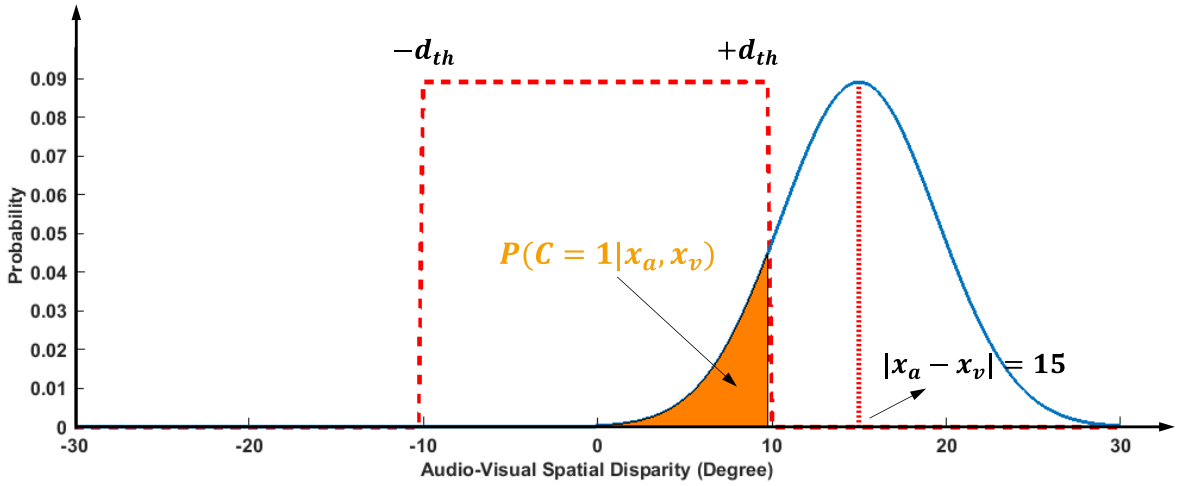


FIGURE 5-7

A depiction of the situations in which the subject can possibly falsely perceives the current causal situation as common-source. This misperception is imposed by noise.

will certainly report the pair of Audio-Visual signals generated by two distinct objects. But, when the sensory signals are polluted with noise, perceived disparity d_{av}will be uncertain. Consequently, there is a non-zero chance that the subject misperceives the current situation and reports a common-cause (FIGURE 5-7). Since spatial disparity is a linear function of sensory signals (x_a and x_v), its variance can be directly calculated as sum of the variance for each sensory signal. Therefore, apart from mean disparity, sensory noise also plays an important role in characterizing the behavioral in this task. By shrinking sensory noise variance, or increasing the spatial disparity, the frequency of false unity perception becomes lower. In other words, these parameters tune the width of intuitive perceptual binding window described in FIGURE 5-7**.** But, it is unclear which neural mechanism this perceptual binding is handled and where the decision is made? We just know that most likely anterior-IPS preserves the estimation derived by causal inference [Rohe & Noppeney 2015].

To understand the underpinning mechanisms of Bayesian Causal Perception, Ma and Rahmati tried to accommodate this process into a Probabilistic Population Code framework [Ma and Rahmati 2013]. As is argued in this article, the derived circuit is not plausible since it requires log operation and Taylor-series expansion, which are not likely present in cortical circuits [Ma and Rahmati 2013]. On the other hand, this model does not resemble the hierarchical motif uncovered by Rohe and Noppeney [Rohe & Noppeney 2015], and thereby cannot successfully reproduce human data. Nevertheless, the general computations that drive the final decision regarding the causal hypothesis look intractable [Körding et.al 2007] [Ma & Rahmati 2013]. In Appendix C, we have reformulated an approximation of common-source posterior probability, as a Gaussian function of spatial disparity d_{av}. This function can successfully incorporate the main

parameters of the behavioral model, i.e. signal reliability, prior probability of common-source and sensory stimuli (FIGURE 5-2). Equations (5-10) shows the derived function that maps perceived disparity into a belief regarding the current causal structure:

$$P(C=1|x_a,x_v)=\frac{a\,Q(d_{av})\,Pcom}{a\,Q(d_{av})Pcom+(1-Pcom)} \tag{5-10}$$

$$Q(d)=\frac{1}{\sqrt{2\pi(\sigma_a^2+\sigma_a^2)}}e^{\left(\frac{-d^2}{2(\sigma_a^2+\sigma_a^2)}\right)} \tag{5-11}$$

In which P_{com} is the prior probability of common-cause hypothesis. To specify decision threshold d_{th} in FIGURE 5-7, one must calculate the root of Log-PR with respect to d_{av}, where the probability of $P(C=1|x_a,x_v)$ is identical to $P(C=2|x_a,x_v)$. For more detail see equations (C-12) and (C-13) in Appendix C.

As is shown in (5-10), to compute the probability of the common-source, the circuit must perform division operation, and a radial-base function, that are both present in neural circuits. The tuning function of almost all place-coding neurons are radial-base, and divisive normalization is usually computed using nonlinear lateral inhibition in cortical circuits [Pouget & Sejnowski 1997]. From another point of view, equations (5-10) and (5-11) are the results of marginalization process (see Appendix C). However, since variables are normally encoded within neural activities (in our model we use PPC), this makes the problem more complicated. Because it is necessary to preserve information content while neural circuit transforms x_a and x_v to d_{av} or equivalently $P(\,d_{av}|D_{av})=P(\,d_{av}|R_a,R_v)$. One important feature of PPC is encoding signal variance σ_i^2 in the gain of amplitude, i.e. $\sigma_i^2\propto\frac{1}{g_i}$. Since $\sigma_d^2=\sigma_v^2+\sigma_a^2$ and D_{av} is the PPC-encoding of d_{av}, one requirement for marginalization circuit is to automatically encode $g_d=\frac{g_a g_v}{g_a+g_v}$.

To fulfill these requirements, we have used a neural model similar to that of introduced in Section 5.2.1, but with different pattern of synaptic weights and without reciprocal connections. FIGURE 5-8 shows the architecture of this network. First, the activity of R_v and R_a are copied into a common frame of reference M, then, they are normalized using divisive normalization. But, the main function of divisive normalization in this case is to achieve an optimal transformation. So, the activity of neuron M_{lm} in a common-frame of reference can be computed as follows:

$$M_{lm}=\frac{r_v^l\,r_a^m}{\Sigma(c_v^k\,r_a^k)+\Sigma(c_v^k\,r_a^k)} \tag{5-12}$$

Where, coefficients $c_v^k=\left(\frac{1}{\sigma_{tc_v}^k}\right)^2$ and $c_a^k=\left(\frac{1}{\sigma_{tc_a}^k}\right)^2$ are the reverse of tuning width of k^{th} neuron in R_v and R_a. Equivalently W_{klm}^a and W_{klm}^v in FIGURE 5-8 can be formulated as follows:

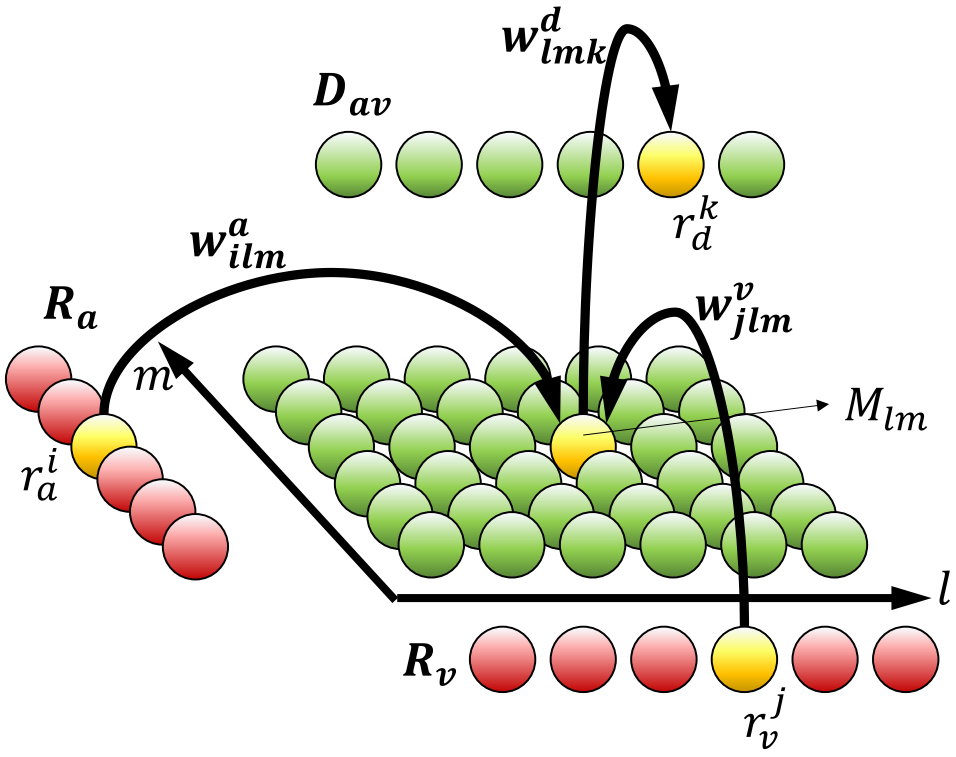


FIGURE 5-8

The architecture of the Marginalization network is depicted. The input populations R_a and R_v are projected into an intermediate neural-sheet M according to W^a_{ilm} and W^v_{jlm}. Each intermediate neuron M_{lm} is connected to disparity-encoding neural population D_{av} in such a way that the network is able to optimally encode audio-visual disparity without information-

$$W^a_{ilm} = \begin{cases} 1 & i = m \\ 0 & i \neq m \end{cases}, \quad W^v_{jlm} = \begin{cases} 1 & j = l \\ 0 & j \neq l \end{cases} \tag{5-13}$$

Now having the activity of intermediate neurons calculated, r^k_d is derived according to a feedforward linear projection:

$$r^k_d = \sum_{lm} W^d_{lmk} M_{lm} \tag{5-14}$$

We have chosen W^d_{lmk} in such a way that the synaptic strength between M_{lm} and r^k_d reflects the neural correlation:

$$W^d_{lmk} = \begin{cases} 1 & |S^v_{tc}(l) - S^a_{tc}(m)| \leq S^d_{tc}(k) + \varepsilon \\ 0 & |S^v_{tc}(l) - S^a_{tc}(m)| > S^d_{tc}(k) + \varepsilon \end{cases} \tag{5-15}$$

Where $S^i_{tc}(j)$ is the tuning center of the i^{th} neuron in the population $j = \{a, v, d\}$, and ε defines the range of excitation. This pattern of synaptic strength is in fact a result of the Hebbian associative learning. The neural framework of FIGURE 5-8 will be optimal under three assumptions: Gaussian noise, linear transformation, and the pattern of synaptic weight mention in (5-13) and (5-15) [Beck et.al 2011].

Having the disparity optimally encoded in D_{av}, the input current of decision neuron C is determined by a linear synaptic projection from D_{av} neurons. Therefore, it includes the variability of sensory signals in causal decision. The activation function of neuron C can be either soft-threshold or sharp-threshold which corresponds to Model-Averaging and Model-Selection respectively. The best candidate for sharp-threshold is the decision threshold where the posterior ratio is equal to unity (see Appendix C). When the posterior ratio becomes greater than one, neuron C will strongly inhibit the segregation pathway and will potentiate the fusion pathway implying that the observed sensory attributes correspond to a single object. This decision strategy is known as Model-Selection [Wozny et.al 2010]. On the other hand, if we chose the posterior probability of equation (5-10) as the activation function of neuron C, the shunt inhibition will be soft and thus the output of Segregation and Fusion pathways can be combined according to the perceptual belief in the current causal situation. This decision strategy is known as Model-Averaging. It is still unclear which decision strategy is employed by human. It is observed that some human subjects perform Model-Selection, and some tends to do Model-Averaging [Rohe and Noppeney 2015].

5.4. Experimental Results

To evaluate the performance of the proposed neural model, we have simulated the spatial ventriloquist experiment performed by Körding and colleagues [Körding et.al 2017]. In a single trial of this experiment the subject is presented by a synchronous audio-visual signal originating from five possible locations along azimuth (see FIGURE 5-9). In [Körding et.al 2017] the visual cue is a high contrast Gabor wavelet extends by 2° on a background of visual noise. Rohe and Noppeney chose a cloud of white dotes on a black screen as visual signals [Rohe and Noppeney 2015]. In both studies a brief burst of white-noise is used as acoustic signal which is presented through a pair of headphones. The duration of both stimuli in [Körding et.al 2017] is set to 35ms and the subject should report the location of acoustic and visual signals using two sets of push-buttons. Each set is composed of five keys associated with five possible locations of stimuli. This experimental paradigm is known as dual-report ventriloquist paradigm and its main purpose is to study the joint audio-visual percept of the subjects [Wallace et.al 2004] [Shams et.al 2005]. In FIGURE 5-9, the schematic representation of this experimental paradigm is shown. Nevertheless, Rohe and Noppeney performed a task-relevance experiment in which the subject should report the position of one of the signals at each trial as well as the perceived causal situation. This is known as task-relevance experimental paradigm [Rohe & Noppeney 2015].

Körding & colleagues trained the Bayesian generative model of FIGURE 5-2 using experimental data collected from 19 subjects. Once the parameters of the causal model

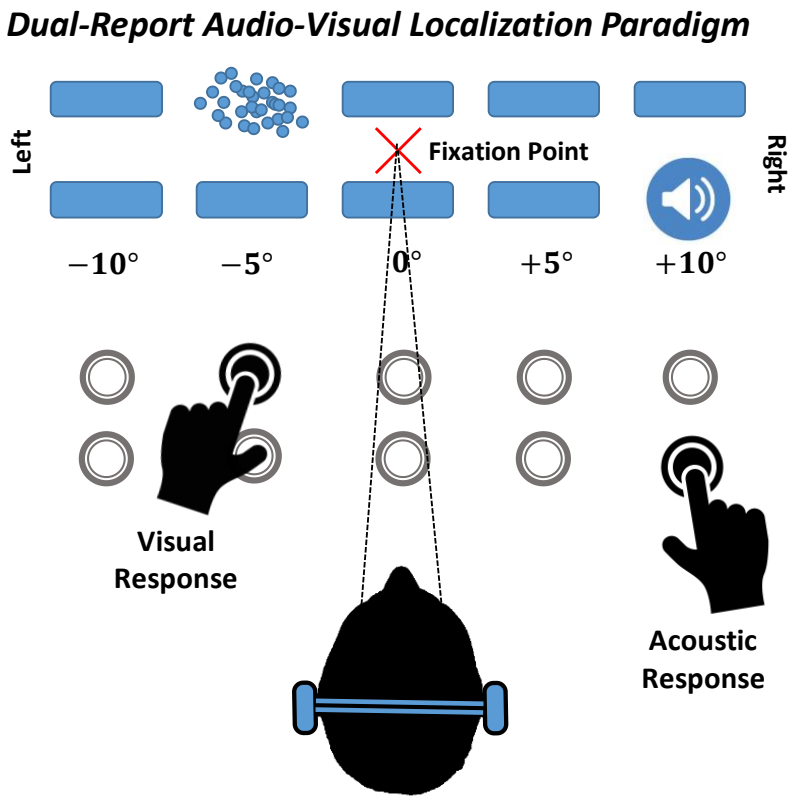


FIGURE 5-9

Schematic representation of dual-report spatial ventriloquist paradigm for Audio-Visual localization task. The subject is presented by a cloud of dots and a synchronous 35ms long burst of white noise, with varying spatial disparity from one trial to another. The location of each cue is uniformly drawn from five possible choices: $\{-10°, -5°, 0°, +5°, +10°\}$. The perceived position of both acoustic and visual signals are reported using two sets of push-buttons. Each set includes five keys associated with five possible locations of either acoustic or visual signal [Körding et.al 2007].

are tuned, they perform a Monte-Carlo simulation with 10000 samples to compare the performance of the trained model with psychophysical data. Given the parameters of the trained model reported in [Körding et.al 2007], i.e. $\sigma_a, \sigma_v, \sigma_p, p_{com}$, we have generated 10000 data samples, and then, tuned the parameters of the reformulated causal model of Appendix C, i.e. $\sigma_a, \sigma_v, a, p_{com}$, using maximum likelihood estimation [Myung 2003]. Thereafter, we have set the parameters of the proposed Causal neural model (FIGURE 5-5) according to this parameter set. In TABLE 5-1, the relative log-likelihood of the fitted model and its parameter set are listed.

5.4.1. Perceived Spatial Unity

The unique feature of Multisensory Causal Inference is to incorporate the belief in the current causal hypothesis in the final estimate. To evaluate the role of disparity in shaping

TABLE 5-1 Parameters of reformulated Causal model and generative Bayesian model.

Model Parameters	P_{com}	σ_p	σ_a	σ_v	a	*Log-Likelihood*
[Koerding et.al 2007]	0.28	12.3	9.2	2.14	75	0
Proposed Model (Appendix C)	0.30	N.A.	8.28	2.51	N.A.	-5.3

the causal perception, the rate of the perceived common-source hypothesis (referred as spatial unity-report in literature), as a function of spatial disparity $S_a - S_v$ is analyzed and demonstrated in FIGURE 5-10. As is shown in this figure, shorter the audio-visual disparity becomes, more often the spatial unity is reported by the subject (dashed red line). However, due to intrinsic uncertainty, even when the signals are perfectly aligned there is still a fraction reports as non-unity scenario. As the disparity becomes wider, it is easier for the observer to the segregate the signals into separate ones and therefore the rate of unity-report decreases. Similar to Mont-Carlo simulation that Körding and colleagues performed, I have also simulated the response of the proposed neural model to 10000 pairs of audio-visual signals, i.e. x_a, x_v. The input signals are generated using Bayesian Generative model of FIGURE 5-2. If the activity of Neuron C in the marginalization pathway which encodes the posterior probability of common-source, exceeds the

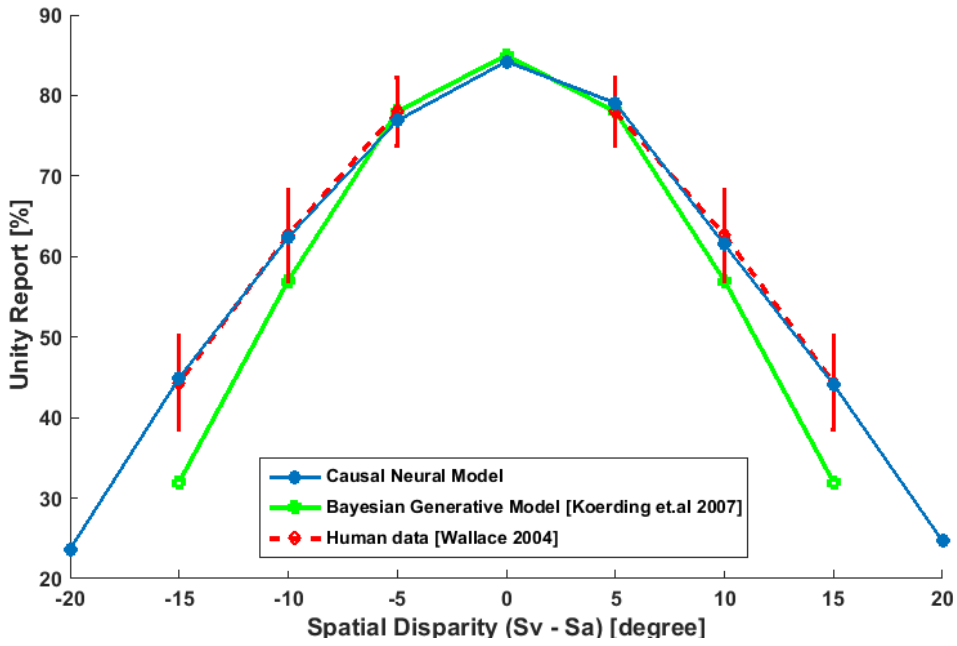


FIGURE 5-10

Spatial-Unity-Report as a function of spatial-disparity is illustrated. The result for the proposed neural model is indicated by blue solid line. The performance of the Bayesian Generative Model proposed by Körding and colleagues is shown by green solid line, and the average response of the subjects reported in [Wallace et.al 2004] is shown by dashed red line.

threshold, i.e. 0.5, the neural observer reports a common-source event (or equivalently spatial unity-report). Once the position of acoustic and visual signals are estimated by the neural observer, one of the five possible choices of location with the closest value is picked up as the final response of the model. The spatial unity-report of the network is depicted by a blue solid line in FIGURE 5-10. As is illustrated in this figure, the neural model can remarkably produce the average behavior of human subjects. It is important to note that Körding and colleagues used the data reported in [Wallace et.al 2004] as a baseline to evaluate the performance of the generative model in replication of human data.

5.4.2. Localization Bias

In multisensory research, bias is commonly referred as a signature of cross-modal interactions [Shams 2012]. The ventriloquist effect of more reliable signal in sensor fusion is a well-known example of perceptual bias. Since in spatial perception, vision is usually the dominant modality, the perceptual bias for acoustic signal as a function of disparity is commonly evaluated in literature [Wallace et.al 2004] [Roach 2006] [Körding et.al 2007] [Sato et.al 2007]. This criterion is formulated according to the following equation:

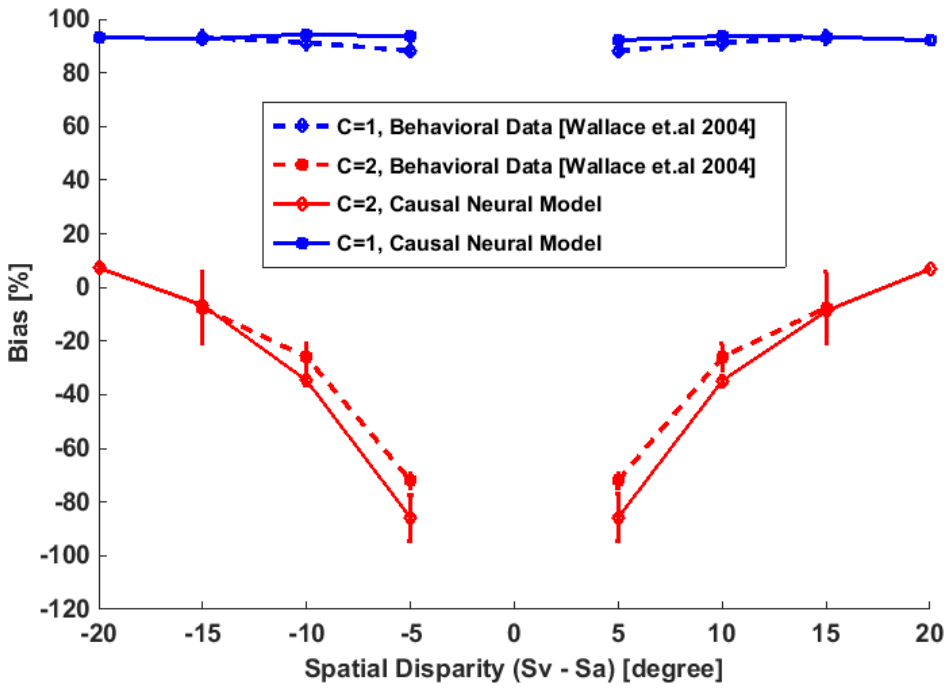


FIGURE 5-11

The mean perceptual bias in the estimated location of acoustic signals are depicted for common-source cases (blue) and independent-source cases (red). The results of the proposed network is indicated by solid lines, and the psychophysical data is shown by dashed lines; with permission from [Wallace et.al 2004].

$$Bias\ [\%] = 100\frac{\hat{S}_a - S_a}{S_v - S_a} \tag{5-16}$$

Where, $\hat{S}_a$ is the average acoustic response of the observer over the trials with spatial disparity equal to $S_v - S_a$.

In FIGURE 5-11 it is examined how the perceived causality influences the estimated location of the acoustic signal according to (5-16). The dashed-lines represent the psychophysical data reported in [Wallace et.al 2004] and the solid lines show the average acoustic bias derived from proposed neural model and evaluated for 10000 audio-visual signals. As is depicted in FIGURE 5-11, when a common-cause is perceived by observers - either human observers or causal neural model, the acoustic location is strongly drifted towards the position of visual signal. As a result, the average bias is so high in this case (blue line in FIGURE 5-11). On the other hand, when the subjects perceive the signals as events caused by distinct sources, the acoustic signal is perceived away from the original location of the stimulus thus that generates a negative bias (red curves in FIGURE 5-11). This counterintuitive phenomenon cannot be predicted by classical models of multisensory integration [Alias & Burr 2004a] [Ernst & Di Luca 2011]. As is illustrated in FIGURE 5-11, the proposed causal neural model is also capable of capturing this specific cross-modal characteristic.

The effect of negative bias rapidly vanishes as the spatial disparity becomes wider. Körding and colleagues argued that this is a selection-driven bias originates from the fact that we calculate the bias exclusively for trials that are perceived as non-unity cases [Körding et.al 2007]. As a result, the distribution of acoustic responses within these trials is a truncated Gaussian distribution (similar to the colored area in FIGURE 5-7) in which a part of Gaussian profile corresponds to common-cause is truncated away. Eventually, that leads to a negative bias because the mean of truncated Gaussian is skewed from the center of Gaussian. As the spatial discordance becomes wider, the mean of truncated Gaussian moves away from discrimination threshold (see FIGURE 5-7), and thus the truncated Gaussian becomes smaller and thus the negative bias vanishes.

5.4.3. Motor Confidence

One important criterion that reflects the uncertainty of the responses in each causal situation is *motor confidence*. Motor confidence is defined as the standard deviation of responses to audio-visual stimuli. We have plotted the values of this criteria as a function of disparity for non-unity cases. Interestingly, the average confidence of the subjects to choose the perfectly congruent signals, i.e. $S_a - S_v = 0$, as non-unity case is too low (or equivalently the standard deviation of motor response is too high). This implies that the subjects are in fact not confident about the wrong choice they made. As is depicted in FIGURE 5-12, by increasing spatial disparity, the confidence of motor action also increases. This means for disparate signals, it is easier to recognize them as signals generated by

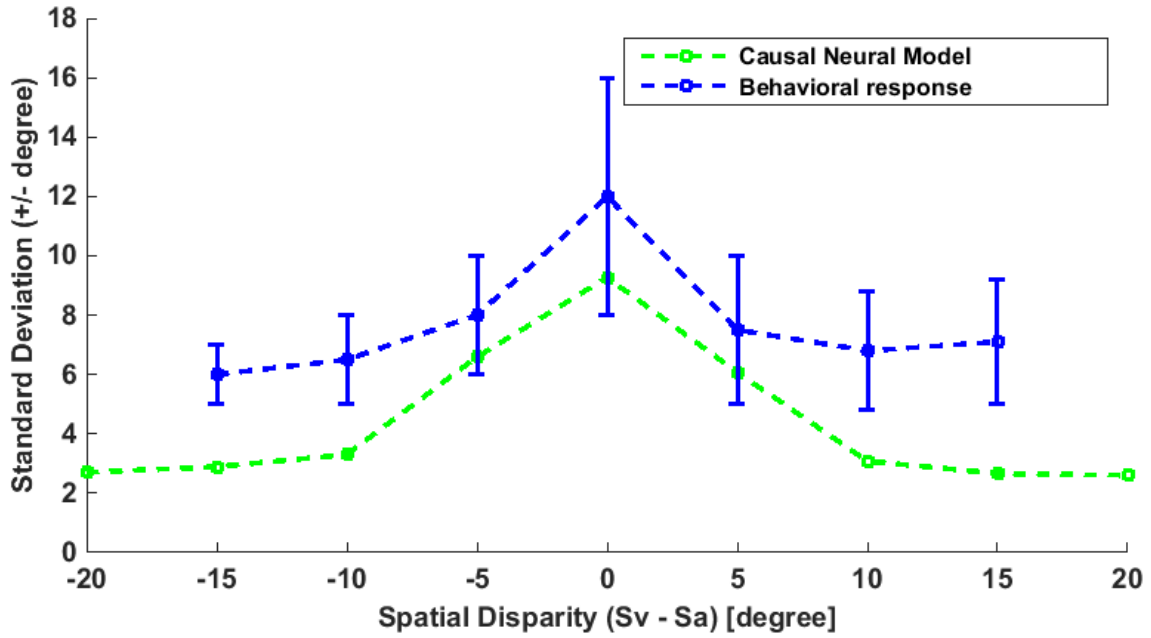


FIGURE 5-12

The motor confidence as a function of spatial-disparity is depicted for independent-sources scenarios: the response of causal neural model is plotted by green dashed line, and psychophysical data is reprinted from [Wallace et.al 2004] with permission (blue line).

independent-sources thus the confidence becomes higher. The neural model can follow the general profile of motor confidence of human subjects, even though it is not perfectly identical to it.

5.4.4. Parameter Sensitivity in Causal Neural Model

Synonymous to the generative model of Causal Inference, the proposed neural model also consists of four general parameters: sensory noise variance σ_a^2 and σ_v^2, range of sensory observation $\left[+\frac{a}{2}, -\frac{a}{2}\right]$, and prior probability of common-cause p_{com}. Sensory noise is internalized within the gain of neural populations in early sensory cortices, i.e. g_a and g_v in FIGURE 5-5 and FIGURE 5-4. The range of sensory observation and the prior directly influence the posterior ratio of causal hypothesis and thus are incorporated in the marginalization pathway, where the firing activation function of the decision neuron C forms the posterior probability. However, as is discussed in 5.3.2.2, sensory noise also contributes in perceptual decision. This contribution is implicitly incorporated within the encoded variability of the spatial disparity in D_{av} which determines the input current of the decision neuron C (see FIGURE 5-5). In FIGURE 5-13–Left we have analytically evaluated the sensitivity of causal model to sensory noise. Regardless of the standard deviation of visual signals, the decision threshold is monastically increasing as the prior probability of the common-source hypothesis p_{com} rises. This implies that, the observer tends to bind

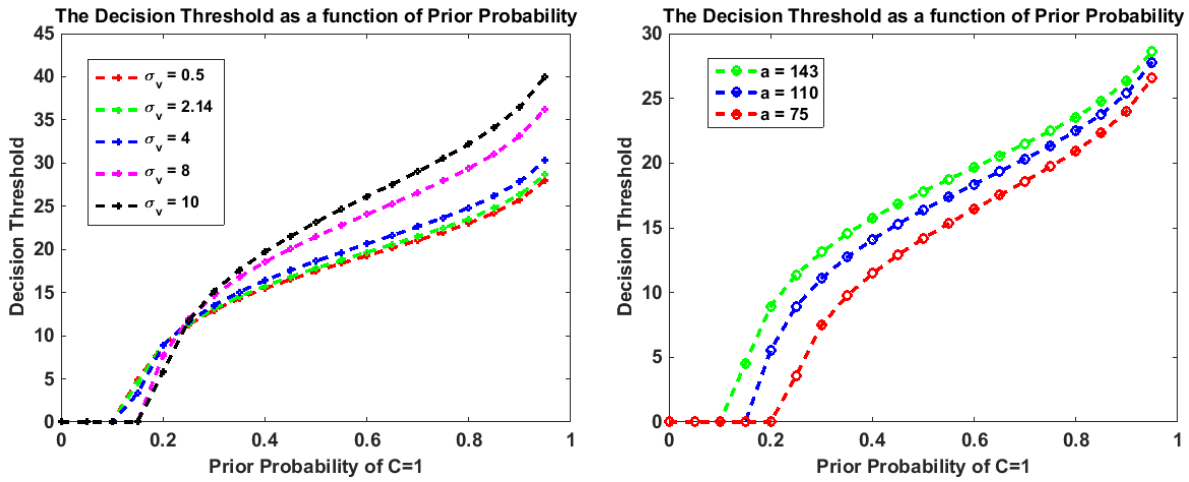


FIGURE 5-13

Decision Threshold as a function of Prior Probability of Common-Source hypothesis is depicted and analyzed for different sensory noise values (Left), and different range of sensory observation (Right). In Left diagram $\sigma_a = 8.3$, and $\frac{a}{2} = 70°$. In right figure $\sigma_a = 8.3$ and $\sigma_v = 2.5$.

signals into a single estimate more frequently. Moreover, for a constant value of prior, higher sensory uncertainty leads to a wider decision threshold because that widens the width of posterior ratio (see equation (C-10)). In fact, when sensory noise is high, the observer cannot easily discriminate the disparate signals and thereby the frequency of the common-source reports becomes higher. This is reflected by an increased decision threshold. As is depicted in FIGURE 5-13–Right, the range of sensory observation a modulates the decision threshold as a function of prior. Since a reflects the prior assumption regarding the possible sensory signals, higher it becomes, more likely the signals originate from separate sources.

5.5. Remarks

The process of causal inference in the context of multisensory perception is discussed in this chapter. During this process the observer should first compute the belief in the possible cause that generates current sensory signals, to decide whether to combine them into a single estimate or segregate them. This form of computation is situated in a higher level of complexity compared with forced-fusion. From functional point of view, typically higher cortical areas generate more complex process along a hierarchy. A recent fMRI study revealed that the process of multisensory causal inference is performed within a distributed hierarchy in cortex. The results of this work uncovered the cortical regions that instantiates specific components of perceptual causal inference in Audio-Visual localization. Segregation is programmed in early sensory cortices, forced-fusion is

performed by posterior parts of IPS[53], and causal estimate is preserved within anterior parts of IPS. This hierarchical representation is compatible with the hierarchical Bayesian model that can remarkably describe the psychophysical data. However, the underlying neural mechanism that mechanistically generates this process is still questioned.

Previous neural models of multisensory causal perception either suffer from implausibility or cannot fully reproduce the characteristics of multisensory causal inference. To address this problem and having a basic architecture of the components of this process revealed, we have posed three essential questions: the problem of optimal encoding, optimal fusion, and marginalization. In Section 5.3.1, I have proposed a plausible probabilistic neural coding that can preserve the information content of a Gaussian-like random variable within a population of Poisson neurons. The encoding populations in fact provide the input of the neural model. In Section 5.3.2.1, a forced-fusion circuit is proposed, and it is demonstrated that this model can optimally combine the encoded variables into a single estimate. The reliability of each sensory signal is incorporated in this circuit by modulating the gain of the corresponding population activity according to the reliability. To compute the probability of the causal hypothesis, we argued that human observer utilizes the cross-modal incongruence as an intermediate signal. Accordingly, under reasonable assumptions we have reformulated the generative Bayesian model of Causal Inference in such a way that the posterior probability of common-cause can be approximated as a function of spatial disparity (see Appendix C). This function includes computational components that exist in cortical circuits, i.e. reference alignment, divisive normalization, and exponential radial-base function. In Section 5.3.2.2, we have proposed a circuit that can compute the posterior probability of causal hypothesis, given the probabilistic neural populations. One important requirement that is fulfilled in this framework is preserving the variability of the encoded signals in neural transformation. Once the posterior probability is computed, the neural projections of full-fusion and full-segregation pathways to the anterior-IPS are modulated using shunt inhibition. Strong shunt inhibition can implement a model-selection decision strategy, while soft-inhibition can perform model-averaging. The prior probability and the range of sensory observation are internalized in this pathway.

The results and simulations show that the proposed neural model can remarkably reproduce the main characteristics of perceptual causal inference in an audio-visual localization experiment. This work is the first neural model that can successfully replicate the complex process of multisensory causal inference using realistic neuro-computational principles. We have presented a mechanistic way of how possibly cortical hierarchy performs the process of multisensory integration, and specifically how the statistical parameters of a sensory space can be internalized within a neural circuit. The proposed

[53] Intraparietal Sulcus

model also predicts the role of sensory likelihood, spatial discrepancy, and importantly, the prior probability of the causal structure in the spatial binding window. Basically, less uncertainty in the sensory likelihood shrinks the binding window or equally sensitizes the subject to discriminate common-source or separate-sources hypothesis.

Using a mechanistic model-based approach, in this work I have specifically evaluated the role of sensory noise, prior probability, and particularly the spatial incongruency in guiding multisensory causal perception. The less focused factor in this process is the role of temporal congruency. One interesting future work is to use the dynamical Bayesian model to test how temporal incongruency shapes the behavior during multisensory integration. In addition to that, the role of neural synchrony between intraparietal sulcus and early sensory regions, can be analyzed as another future work. I believe that the functional correlation between the computational components of hierarchical causal inference and the specific regions of the cortical hierarchy is likely produced by a synchronization mechanism. We think this neural synchronization might be a key in binding multisensory signals in time and space simultaneously. This requires a more detailed neural model that includes temporal coding. A recent study shed some lights on this hypothesis [Keil and Senkowski 2018].

Chapter 6

Summary and Discussion

"From the cradle to the grave, seek for knowledge"

*— **Prophet Muhammad**, blessings of Allah be upon him and his family.*

For the last decades, Helmholtzian notion of Perception has been a dominant theory in Brain Computing that shapes a wide range of algorithms in Machine Learning. After decades, there is still no convincing reason to deny his view on brain computing [Kiefer 2017]. In this theory, perception is defined as a subjective inference process by which the observer computes an internal belief in the possible hypotheses regarding the state of the world, given the sensory evidence of the state variables. Probability theory and specifically Bayesian Decision Theory has bestowed a systematic methodology to combine a priori belief (prior probability) with the evidential information (sensory likelihood), in order to compute the posterior probability of the state variable. Although, cue integration can take place within the scope of a single modality, perception is mostly multimodal since the sensory events are mostly composed of multiple attributes. On the other hand, perception benefits from multisensory integration in many ways: minimizing the imprecision, sensory recalibration, perceptual learning, and brain development. This process is mainly handled by a hierarchical processing that represents different aspects of the world within different levels of complexity.

The hierarchy in functionality and data structure (e.g. simple visual features vs semantic features), is reflected by a hierarchical architecture in circuit-level. Generally, higher the order of cortical circuit, implies a more complex form of data and processing. For instance, in visual pathway, *V1* and *V2* collect visual information from thalamus to create a simple feature-map in retinal-coordinate while V5 consist of more complex neurons that compute a rough pattern of visual-motion. Thereafter, MSTd neurons receive this motion information and combine it with vestibular and ocular signals to perform a more complex form of computation in head-centered coordinate. The process of aligning data in different coordinate systems is known as *reference alignment* in multisensory integration. The hierarchical processing in the sensory cortex is an inevitable principle that is computationally advantageous and leads to a coherent form of perception.

In Chapter 2 we summarized the probabilistic and deterministic methodologies that have formalized the Helmholtz theory in the context of multisensory perception. Rationality is a key characteristic of the perceptual system that is equivalent to coherency-maximization [Kiefer 2017]. A rational observer will not fuse distinct or inconsistent sensory signals into a single estimate. In this case the rational observer either penalizes the factor of contribution for the inconsistent sensory node or recalibrates it according to the perceived value. In deterministic approaches the coherency function is defined by a deterministic relation function. This function relates the connected sensory nodes and thus determines whether they produce plausible signals with respect to each other, i.e. is a sensory signal comparable with the value predicted by other nodes? On the other hand, one benefit of this form of integration is that the observer can predict the expected value of a dropped sensory node.

Voting-based algorithms, Democratic-Integration and Mutual-Prediction technique are amongst these methods (see Section 2.2.1 and 2.2.2). One benefit of such approach is demonstrated in FIGURE 2-4 where a distributed network of sensory nodes is represented. The sensory nodes are connected to each other according to specific functions (see FIGURE 2-3) that reflects the physics of the signals. Given the instantaneous values of the sensory signals including retinal action potentials and inertial movement of the camera, the estimated value of the intermediate variables will be updated in such a way that all connected nodes agree according to the relation functions. Eventually after few iterations, the network will reach a relaxing state where the light intensity for each pixel is computed. Note that there is no sensor to measure the light intensity and it is estimated by means of retinal action potentials and inertial motion of the camera. The dynamic of this network is called *relation-satisfaction* in the literature. Moreover, this network (see FIGURE 2-3) is inspired by the theory of distributed cortical responses depicted in FIGURE 1-5 but in a deterministic regime.

In FIGURE 4-15, a neural model of a *relation satisfaction* network is proposed. The proposed model can first learn the relation as a function between one of the connected variables and other variables (as an independent variable). Rather than interacting sensory nodes there are interconnected neural populations that each encodes a single variable. The relaxing dynamics is also implemented using a plausible neural dynamic which is called Attractor Dynamics. The attractor surface is in fact a multidimensional hyperplane in which the encoded variables relax into one single point of that. The modification of more reliable signal will be smaller than less reliable signal. The reliability is encoded in gain of neural activity. As is demonstrated in FIGURE 4-17, an experiment is simulated in which one population is initialized with two bumps of activities, the one which is not in agreement with two other sensory values is totally whipped out, and the coherent one is potentiated.

In FIGURE 4-19, I tested a nonlinear scenario in which a quadratic non-one-to-one relation function is trained. As is shown in this figure, the attractor dynamic can restore the value of dropped sensory node according to the trained relation and the initialized values for other sensory nodes. The coordinate system of the sensory modalities are sometimes related according to a specific function. In this case the relative flexibility of the proposed network of FIGURE 2-15 allows us to perform reference alignment. Once the network is trained the relation between sensory cues, this function can align the sensory cues accordingly. However, the stability of network's dynamic for complex functions is not guaranteed. There is a chance the network becomes unstable particularly when the surface gradient is too large.

In probabilistic methods, a rational strategy for information integration is to minimize the Mean Square Error between the value of the perceived signal and the physical stimulus (or sensory feedback). This leads to the problem of *optimality* in sensor fusion. An optimal algorithm should combine the sensory attributes in such a way that MSE becomes minimum. The quality of a sensory attribute can be indicted by a factor which shows how error-prone that node is. This is known as *validity problem* in sensor fusion. In Section 2.3.1, it is mathematically proved how MSE minimization leads to an optimal linear combination of the sensory signals (*Maximum-Likelihood-Estimation*). In this algorithm the weight of combination for a single node (the quality of the signal), is reversely proportional to signal variance.

MLE is a basic fusion algorithm that can account for the problem of *validity* and *optimality*. However, the optimality of MLE is guaranteed under specific constraints i.e. noise process is assumed to be additive-Gaussian, sensory nodes are statistically independent, and the prior probability is assumed uniform. To generalize MLE and to incorporate the prior distribution of sensory observation, Bayesian Integrating algorithm is introduced (see Section 2.3.2). FIGURE 2-6 illustrates the main benefit of incorporating the prior in Bayesian integration model. As is shown in this picture, the variance of posterior probability of the perceived signal is reduced as compared to MLE and prior probability. However, it leads to a perceptual bias by which the final estimate is drifted toward the mean of prior probability. The error is not exclusively derived by random noise. Sometimes one sensory node is persistently drifted away from the real value of the physical stimulus. In this case the systematic bias will be directly imposed into the final estimate regardless of the signal variance. That is inevitable for both MLE and Naïve Bayesian Integration algorithms to exclude the effect of bias in the final estimate.

This is one of the costs of integration that must be balanced with the benefit of variance minimization. In Section 2.3.3, an extended Bayesian framework is introduced called *Coupling-Prior model* that provides a practical model to identify and to distinguish bias-driven error from noise-driven imprecision. *Coupling-Prior* model integrates multiple

processes of multisensory integration - calibration and remapping, forced fusion, and segregation - in a unified framework [Ernst and Di Luca 2011]. Although this approach provides a powerful model, it is doubted that the perceptual system implements an embodied form of prior that constantly changes from trial to trial [Shams 2012].

Summarizing some psychophysical evidences in human and monkey, in Chapter 2 we show how behavioral data fits to Bayesian Integration and MLE in a broad spectrum of perceptual inference tasks [Ernst & Banks 2002] [Alias & Burr 2004a] [Wallace et.al 2004] [Shams et.al 2005] [Bresciani et.al 2006] [Körding & Wolpert 2006] [Ernst 2007] [Wozny et.al 2008] [Fetsch et.al 2009] [Ursino et.al 2011] [Petzschner & Glasauer 2011][Rohe & Noppeney 2015] [Boyle et.al 2017]. Although, these models capture some aspects of multisensory perception in human behavior, almost all of them left a question open: *what is the neural correlates of MLE and Bayesian Integration in sensory cortex?*

Understanding the neural mechanisms of this Bayesian behavior requires the analysis of neural activities in multisensory areas. There are very few physiological recordings in monkey that explained how the firing rate of some neurons in posterior and intraparietal regions is directly correlated with the value of log-posterior-odds assigned to a set of trained visual cues [Yang and Shadlen 2007]. However, it is unclear whether LIP converts the visual information into probabilistic values, or it is provided by neurons in ventral pathway. Yet, despite these plentiful behavioral evidences, it is not clear how the cortical circuits preserve the probabilistic quantities into a single estimate, and they are combined.

In addition to that, the statistical parameters of a Bayesian Model must be accommodated in a neural circuit, whether implicitly or explicitly. Most of the proposed neural models of cue integration are based on a straightforward hypothesis which suggests that the uni-sensory signals are pooled in a ploy-sensory convergence zone. The convergence zone is the place the cortex creates a unified forms perception. But, this theory seems not realistic since there are many poly-sensory regions in the Brain that are mutually interconnected. Using realistic neuro-computational principles e.g. probabilistic population coding with Poisson variability, Attractor Dynamics, and Gain-Field Modulation, it is explored theoretically how multisensory integration can be performed within a distributed circuit. In FIGURE 4-20 a tri-modal version of the proposed neural model is shown, where the interaction of the hypothetical distributed cortical regions performs sensor fusion. Each region is accompanied with a supplementary circuit that analyzes the statistics of the last encoded values in order to compute signal variance. Then, the variance modulates the synaptic projections of the corresponding uni-sensory neurons to the poly-sensory neurons (see FIGURE 4-20). This mechanism of shunt inhibition is equivalent to gain modulation. As a result, the weight of each population in fusion is proportional to the quality of signal. We perform a tri-modal heading estimation

experiment using a robotic apparatus. The results show this approach is near optimal and almost identical to MLE (see FIGURE 4-21). Intuitively, gain modulation changes the attractor dynamics in favor of more reliable cue. In other words, more reliable a neural population is, higher gain of activity it has and less it will be modified and more it contributes in the unified value of perception. It is argued in Chapter 4 that using optimal sensory coding, this network can be scaled up into a more generalized form of Bayesian Integration, by incorporating the prior probability.

The idea is that the tuning curve of the neurons that encode values close to fovea should be carefully shrunk, and the tuning for the peripheral neurons must be widened. Some theories postulate that cortical-maps employ a similar approach to implicitly describe the statistics of the experienced sensory data. The shrinking factor is called *cortical magnification factor* in *Self-Organizing-Map* network [June 1991]. We suggest that a similar approach can be used in attractor networks to incorporate a priori information of the sensory signals. Analyzing the effect of cortical magnification factor in attractor dynamics can be considered as a future work.

Studying the interplay between multisensory integration and attention in has been the center of the attention lately [Macaluso 2012] [Rohe & Noppeney 2018]. Most of the computational models of visual attention are based on saliency-map (see this survey [Filipe and Alexandre 2015]) or a priority-map [Bisley 2011]; where the most salient stimulus is emerged by a competitive neural process. Rougier and colleagues have proposed a neural model of visual attention in which a dynamic interplay between inhibitory and excitatory synapses determines the location of the salient object [Rougier & Vitay, 2011] [Rougier 2006] (see FIGURE 4-1). This model fails to register the location of the target in complex scenarios, e.g. when the focused object collides with a moving salient distractor. In Chapter 4 of this dissertation, this network is scaled up and a new recurrent hierarchical network is proposed (see FIGURE 4-2). By fusing the predictive location of the target in this network using motion-cue, this hierarchical model can overcome the problem of losing focus in collision site. On top of this hierarchy, and for a single receptive field of attention field, a motion sensitive population of neurons is located that are laterally connected. Motion sensitive neurons receive information from a hidden-layer of neurons. Hidden layer consists of context neurons that preserve a history of the hidden neurons' activity. The hidden neurons are connected to the attention field using a feedforward connection. These synaptic connections are trained using Dynamic Error Back Propagation algorithm to give an estimate of the visual-motion for each receptive fields of the focus map.

Motion-detectors predict the location of the target in next time step and accordingly provide a feedback signal to the focus map. The overlapping neural activity of attention field with predictive pattern of activity causes a stronger neural activity in attention field

and thus it helps the observer to cancel out the colliding distractor. In FIGURE 4-3, we have shown how the overlapping pattern can occur in an example. As we might expect from information theoretic point of view, fusing a new form of information i.e. motion, makes this network much more robust against noise. In FIGURE 4-7 the results of the noise analysis experiment (see FIGURE 4-6) is shown where the error between the location of the captured target and its original location is plotted.

As is demonstrated, the error is drastically reduced in the proposed network as compared to the previous approach. To evaluate the performance of this model in a colliding scenario which is beyond the power of saliency map approach, two experiments are conducted. The first uses artificial data in which two moving bumps of activity, one as a distractor and another as a target, are moving in the field of view in such a way that they collide in near fovea (see FIGURE 4-8). Even though the network observer is distracted for a short time after collision towards the salient distractor, but the motion-cue can reallocate the location of the target to the observer's attention, FIGURE 4-8-b. It is also illustrated that the basic network is not able to handle this task at all (see FIGURE 4-8-a) In a more realistic experiment, the recorded events from a neuromorphic silicon retina sensor are used (see FIGURE 4-11) where two persons are moving in the field of view. The allocated location of attention is marked by a blue rectangle. This task is a difficult scenario for saliency-detection based networks. In addition to these experiments, a real-time robotic experiment is conducted using a 6-DOF robotic-head equipped with a pair of silicon retinas and high precision actuators. The robot is presented by a blinking laser pointer as a target and a big NST letters at the background as distractor. The network is implemented in *CUDA-C* using *nVIDIA*GPU in order to run the network in real-time. In FIGURE 4-10, the results demonstrate the direction of detected motion along with the allocated location of attention. To avoid computational complexity, the proposed network integrates the direction of visual motion, and accordingly modulates the process of attention allocation. To reduce the sensitivity of the network to the velocity of the target, as a future work it is worth to investigate how possibly the velocity can also be integrated within this model.

Dynamic Vision Sensors allow efficient solutions for various visual perception tasks, e.g. surveillance, tracking, and motion detection. The superiority of this kind of sensors includes: the high dynamic range of light sensitivity, reducing the redundant static features, and very fast temporal resolution. Similar to retinal photoreceptors, any perceived light intensity change in the DVS generates a single event at the corresponding pixel. DVS thereby generates a stream of spatiotemporal spikes (events) to encode dynamic visual features [Lichtsteiner et.al 2008]. This form of representing the visual information has created a new paradigm in vision research. However, that calls for developing radically new asynchronous and event-based information processing algorithms.

This issue and particularly the problem of stereo matching in event-based cameras are considered as a big challenge in the literature [Kogler et.al 2011] [Rogister et.al 2012] [Carneiro et.al 2013] [Camuñas-Mesa et.al 2014] [Firouzi & Conradt 2016] [Osswald et.al 2017] [Dikov et.al 2018]. Most of the existing stereo matching algorithms using DVS either are rooted in classical frame-based methods or they exclusively account for temporal correlation [Kogler et.al 2011] [Rogister et.al 2012] [Carneiro et.al 2013] [Camuñas-Mesa et.al 2014]. In order to fully take advantage of DVS sensors, developing an efficient event-driven algorithm is critical. In Chapter 3, I have developed a fully event-based disparity matching algorithm for visual depth perception using a dynamic cooperative neural network (see FIGURE 3-4). The main idea is to fuse incoming events according to two main geometrical and temporal constraints in order to solve the correspondence problem.

Finding the matching objects in stereo images is known as *correspondence problem* in vision. The important cue that is the outcome of solving correspondence problem is retinal disparity: the relative difference in retinal location of a single object in stereo sensors that reflects the depth of the object see FIGURE 3-1 and FIGURE 3-4 (b). The neural dynamics apply two geometrical constraints: cross-disparity uniqueness-constraint and within-disparity continuity constraint, see FIGURE 3-4 (a). The first implies that for an identical single feature (or event) there must be a unique perceived value of disparity. The second constraint is a result of the fact that an object has a cohesive form and thereby should generate a smooth map of retinal disparity. Synonymous to laminar structure of the cortical circuits, the network is composed of layers of disparity-sensitive neurons. Each single cell corresponds to one possible matching between a pair of pixels in left and right hemispheres. Equivalently, a single cell is sensitive to a single retinal disparity which is equal to the difference of two pixels' location (see FIGURE 3-4 (c)).

To implement these constraints in a neural circuit, the cross-disparity uniqueness is realized by two patterns of inhibitions (red-colored cells in FIGURE 3-4 (d)). Within-disparity continuity is implemented by excitatory synapses within each disparity layers, (green-colored cells in FIGURE 3-4 (d)). The cells are leaky to preserve a short history of the previous events. When a single event captures by one of the retinas, it will be fused into the network and will change the activity of cells. Of one single cell wins the competitive process and exceeds the threshold, it annotates the perceived value of disparity and thereafter will suppress the connected cross-disparity cells. Besides, the winner cell potentiates the neighboring cells that lie at the same disparity layer. This cooperative process leads to an asynchronous extraction of the disparity value for the incoming events without any need to frame them in time. We have tested the performance of this network in several experiments; our results demonstrate the outperformance of the event-fusion in contrast to frame-based fusion that generates a considerably smoother disparity map completely event-based, see FIGURE 3-5 and FIGURE 3-6. Even when the scene is composed of temporally-overlapping stimuli, the network dynamics can cancel out mismatching

patterns successfully (FIGURE 3-9). The results in this work show a significant enhancement in the quality of event-based 3D-reconstruction compared with other methods, and therefore placed the proposed approach as one of the first successful attempts to solve the problem of stereoscopic fusion in event-based silicon retains. However, since there is no mechanism in this network to distinguish whether the event occurs in right-side or left-side retina, for some events it leads to self-side matching (mismatching in bottom bar-graph of FIGURE 3-9). In [Dikov et.al 2017], for a single disparity-sensitive neuron we have proposed a supplementary neural circuit which solves this problem by filtering out the events that might cause this problem. Here in this work the viability of event-fusion as opposed to frame-fusion is demonstrated, although a simple feature is used to solve the problem of matching. Stereo matching based on high level features like lines, contours, and objects will enhance the quality of the constructed 3D map. That can be potentially considered as a future work.

The process of perception is context-dependent. Human observer can recognize at which context which strategy should be taken to fulfill the goal, e.g. optimality. The observer can recognize the association/dissociation of the signals by comparing the contextual, spatial, or temporal characteristics of the signals. For instance, it is not rational to integrate the sound of mewing with a picture of a cow since they are not correlated. The location of a bird singing on the tree is different from that of sitting silent next to your window. The process of credit-assignment in sensor fusion is dealing with the question of whether the signals must be integrated in case they are associated or segregated if they are not associated with each other. This process is placed in a higher level of cognition as compared to forced-fusion and intrinsically involves an inference. Given the noisy sensory observation, the observer should form a criterion in time, space, or in a high-level feature space, to measure the congruency of the signals. Thereafter the belief in the sensory setup that generates the current observation must be computed accordingly. Having the present hypothesis inferred, the observer can combine or segregate signals. This process is known as *Multisensory Causal Inference* in perception and is not an easy problem to solve for nervous system, see FIGURE 5-1.

The first research work that differentiates Causal Inference from conventional forced-fusion is done by Wallace and colleagues [Wallace et.al 2004]. Almost all early psychophysical experiment in multisensory perception focused on specific causal situation at which the signals originate from an identical source. Wallace and colleagues studied the characteristics of human behavior in response to audio-visual signals. Within experimental sessions the signals are drawn randomly to be spatially and temporally incongruent or congruent (see FIGURE 5-11). [Wallace et.al. 2004]. They observed a perceptual bias in reporting the location of the acoustic signal which is highly correlated with the value of spatial and temporal congruency. On the other hand, when the subjects report a perfectly congruent situation, they follow forced-fusion and combine the signals

according to signal reliabilities. Moreover, they observed that the perceptual bias is still present even when the subjects report a common-source situation. Then, they imply that this cross-modal pattern of response must be emerged within two distinct pathways [Wallace et.al 2004]. Following this work, Kördingand colleague developed a Hierarchical Bayesian model that remarkable fits to human data [Körding et.al 2007], see FIGURE 5-2. This model provides a theoretical proof of what Wallace and colleagues suggests. In addition to that, they postulate that MCI should be likely handled within a hierarchical circuit in cortex as is functionally hierarchical, but left the question open what is the neural mechanism that generates this process? This notion is in fact one of the main driving hypotheses of this thesis**:** *the principle of decentralized computation in cortical circuits* (see FIGURE 1-5). There are very few attempts to shed some light on understanding the mechanics of multisensory causal perception in sensory cortex. Weisswange and colleagues used machine-learning techniques to train a feedforward radial-base-function network which reproduces some aspects of MCI [Weisswange et.al 2011]. However, this model cannot account for the pattern of perceptual bias reported by Wallace, and it is also not a plausible model. Ma and colleagues used probabilistic population code to reproduce human data in ventriloquism paradigm (see FIGURE 5-9). They argued that the circuit they proposed is not plausible since it needs to compute log operation and Taylor-series expansion, both are not likely present in cortical circuits. On the other hand, the structure of this circuit is not compatible with recent fMRI data [Rohe & Noppeney]. In Chapter 5, we have proposed a distributed hierarchical circuit for MCI that remarkably can reproduce human data. The structure of this neural circuit is compatible with recent evidences that identified the involving cortical regions during Audio-Visual Causal Perception [Rohe & Noppeney]. Moreover, as opposed to [Ma & Rahmati], the computational elements of this circuit is plausible in cortical circuits. The circuit is composed of three types of neural ensembles located in different levels of hierarchy: early sensory areas that preserve the perceived segregated signals, forced-fusion neurons that preserved the estimated stimulus for common-cause hypothesis, and finally the perceived belief in existing causal hypothesis. To have the process of MCI optimally produced, it is necessary to fulfill three requirements in the model:

1. First, a physical random variable particularly with Gaussian-like distribution, must be optimally represented within a pool of pyramidal cells. The pattern of variability in neural activity can be mostly modeled by a Poisson process. This is referred as the *Encoding problem*.
2. Second, forced fusion is one specific case of MCI that should be computed optimally (*Optimal Fusion*).
3. And third, the neural circuit should compute the belief in the existing causal hypothesis, given uncertain sensory observations. Therefore, there must be a distinct pathway that marginalizes out the nuisance parameters in order to

compute the posterior probability of casual hypothesis (*Marginalization problem*). Since this process is functionally more complicated than fusion and it controls the whole process, thus it must be located at the top of hierarchy.

The first requirement is handled by an optimal linear encoding scheme introduced in Section 5.3.1. This model is known as Probabilistic Population Code that can plausibly represent an arbitrary Gaussian-like random variable within a population of Poisson neurons [Ma et.al 2006]. The main advantage of this encoding algorithm is incorporating the signal reliability in the amplitude of the neural activity. In FIGURE 5-4, we have performed a Monte-Carlo simulation to demonstrate the optimality of this model. As is shown in this figure, the distribution of the decoded variable is Gaussian like. The reverse of the profile's width is linearly proportional to the amplitude of the neural activity. This leads to an optimal circuit for forced-fusion in which a linear combination of two PPCs preserves the combined estimate of two noisy signals. The optimal fusion can be achieved by modulating the amplitude of neural activities according to the relative reliability of each signal, see FIGURE 5-6-(a). A Monte-Carlo simulation is performed to test the optimality of this model. Ten thousand pairs of random signals are generated, encoded in PPC whose gain of activity is modulated according to the reverse of signal variances, and finally the fused estimate is measured by decoding the neural activity of the multisensory convergence zone (multimodal neural population), see FIGURE 5-6-(b). In FIGURE 5-6-(c) the result of MC simulation is illustrated where the distribution of network's outcome is compared with the Posterior distribution of the combined signals. It is demonstrated that they are almost identical. This reflects the optimality of the forced-fusion circuit. However, it is assumed that the noise process in each sensory modality is independent and the sensory observation is uniformly distributed. This circuit provides the information under common-cause circumstances.

To compute the belief in current hypothesis, given the neural activity of early sensory areas, the generative Bayesian model of figure 5-2 is reformulated in such a way that the posterior probability of common-cause can be approximated as a function of spatial disparity (see Appendix C). This function performs marginalization and includes computational components that exist in cortical circuits, i.e. reference alignment, divisive normalization, and exponential radial-base function. The proposed neural circuit of marginalization pathway computes the relative *inconsistency* between encoded variables i.e. visual spatial disparity, see FIGURE 5-8. One requirement for this circuit is to compute the spatial disparity within a neural population in an optimal way.

That means the distribution of the decoded disparity must be comparable with or equal to its posterior probability. In Section 5.3.2.2, it is mathematically proved how to hand-craft the synaptic weights of the network in such a way that the neural population of disparity optimally represents the value of audio-visual disparity. Finally, a linear

combination of the neural activity of this population determines the probability of the causal hypothesis. This probability is encoded by the activity of the decision neuron *C* (see FIGURE 5-4). In the sequel, neuron *C* modulates the synaptic projections of early sensory regions (Segregation) and Fusion pathway to the read-out neurons (motor neuron). The pattern of shunt inhibition can be either soft or hard inhibition which leads to Model-Averaging or Model-Selection decision process respectively.

To test the performance of the proposed model, the hierarchical Bayesian model of FIGURE 5-2 is used to generate sensory stimuli. These stimuli are fed into the network and the results are compared with that of reported in [Wallace et.al] and [Koerding et.al]. As is depicted in FIGURE 5-10, the profile of the perceived unity, averaged for all subjects is remarkably reproduced by the proposed Network. As is shown in this figure, this criterion which is in fact the subjective probability of the common-cause hypothesis, is a function of spatial disparity. As disparity increases, it is easier for the subject to distinguish the signals, and thus the reported rate is decreased exponentially. Another criterion that is known as an exclusive hallmark of MCI is negative perceptual acoustic bias (red curve in FIGURE 5-11). Almost all neural models of multisensory integration are not able to predict this pattern of behavior [Ursino et.al 2014]. The effect of negative bias is because the distribution of acoustic responses within non-unity trials is a truncated Gaussian in which the part of Gaussian profile corresponds to common-cause is truncated away (see FIGURE 5-7).

As a result, the mean of truncated Gaussian is shifted away from the center of Gaussian and that leads to a negative bias. As the spatial discordance becomes wider, the mean of truncated Gaussian moves away from the discrimination threshold. Therefore, the truncated part of Gaussian profile becomes smaller and the negative bias exponentially decreases. The proposed model can successfully capture this characteristic as is demonstrated in FIGURE 5-11. One prominent feature of the proposed model in internalizing the statistical parameters of the sensory world within neural pathways. The sensory likelihood is seemingly implemented within neural activity of sensory specific areas [Simoncelli 2009]. Similarly, the reliability of the signals is incorporated in early sensory populations in proposed mode. The range of sensory observation and prior probability of the causal structure are also internalized in marginalization pathway. These two parameters directly shape the decision of the subject, thus are plausibly internalized in decision pathway.

Appendix A

Linear System Analysis of Coupling-Prior Model Fusion

In Section 2.3.3 of this dissertation, we have described a model of Multisensory Integration which is called Coupling-Prior model, in order to optimally balance the benefit and cost of fusion. In this Appendix, we will derive a linear system description of this model. The outcome of this model is to compute the Maximum-A-Posterior as a linear combination of the partially-reliable sensory signals. The assumptions of the problem are:

- The noise process is additive and Gaussian.
- Noise for each single node, is statistically independent from other nodes and the noise variance is equal to σ_i^2.
- Prior joint distribution is Gaussian-like, and its respective variance is equal to σ_x^2.

For sake of simplicity, here we assume a dual-modal integration scenario. However, it can be scaled up for a multiple cue integration problem.

Two noisy measurements: z_i and z_j, are fetched by perceptual system from physical stimulus $S_w = (S_w{}^i, S_w{}^j)$. Let $S = (S_i, S_j)$ be the sensory signals that might be possibly biased with respect to S_w, so $S = (S_w{}^i + B_i,\ S_w{}^j + B_j)$. Having the problem assumptions, system parameters and variables defined, the sensory likelihood and coupling-prior joint distributions are as follows:

$$P(z_i, z_j | S_i, S_j) = N(S^{MLE}, \Sigma_{MLE}), \Sigma^{MLE} = \begin{bmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_j^2 \end{bmatrix}, S^{MLE} = (S_i, S_j) \tag{A-1}$$

$$P(S_i, S_j) = N(S^P, \Pi), \Pi = R^T \begin{pmatrix} \sigma_m^2 & 0 \\ 0 & \sigma_x^2 \end{pmatrix} R,\ R = \begin{pmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{pmatrix} \tag{A-2}$$

$$\Pi = R^T \begin{pmatrix} \sigma_m^2 & 0 \\ 0 & \sigma_x^2 \end{pmatrix} R = \frac{1}{2} \begin{pmatrix} \sigma_m^2 + \sigma_x^2 & \sigma_m^2 - \sigma_x^2 \\ \sigma_m^2 - \sigma_x^2 & \sigma_m^2 + \sigma_x^2 \end{pmatrix} \tag{A-3}$$

The posterior distribution[54] can be derived by Bayes rule:

$$P(S_i, S_j|z_i, z_j) = P(z_i, z_j|S_i, S_j)\, P(S_i, S_j) \tag{A-4}$$

The product of two Gaussian distributions with covariance Σ^{MLE} and Π, and mean $\mathrm{S}^{MLE} = (S_i, S_j)$ and S^{P}, is the following Gaussian:

$$P(S_i, S_j|z_i, z_j) = N(\mathrm{S}^{MAP}, \Sigma_{MAP}) \tag{A-5}$$

$$\Sigma_{MAP} = [\Sigma_{MLE}^{-1} + \Pi^{-1}]^{-1} \tag{A-6}$$

$$\mathrm{S}^{MAP} = W_{MLE}\, \mathrm{S}^{MLE} + W_P\, \mathrm{S}^{P} = \Sigma_{MAP}\Sigma_{MLE}^{-1}\, \mathrm{S}^{MLE} + \Sigma_{MAP}\Pi^{-1}\, \mathrm{S}^{P} \tag{A-7}$$

To compute the right-hand-side of (A-7), first we should derive W_{MLE} and W_P. But, for that we need to obtain the intermediate matrices including $\Sigma_{MLE}^{-1}, \Pi^{-1}$ and then, Σ_{MAP} :

$$\Sigma_{MLE}^{-1} = \begin{bmatrix} \frac{1}{\sigma_i^2} & 0 \\ 0 & \frac{1}{\sigma_j^2} \end{bmatrix} \tag{A-8}$$

$$\Pi^{-1} = \frac{1}{2\sigma_x^2\sigma_m^2}\begin{bmatrix} \sigma_x^2+\sigma_m^2 & \sigma_x^2-\sigma_m^2 \\ \sigma_x^2-\sigma_m^2 & \sigma_x^2+\sigma_m^2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \frac{1}{\sigma_m^2}+\frac{1}{\sigma_x^2} & \frac{1}{\sigma_m^2}-\frac{1}{\sigma_x^2} \\ \frac{1}{\sigma_m^2}-\frac{1}{\sigma_x^2} & \frac{1}{\sigma_m^2}+\frac{1}{\sigma_x^2} \end{bmatrix} \tag{A-9}$$

Assuming $\sigma_m^2 \gg \sigma_x^2$, equation (A-9) can be simplified as equation (A-10). Then, substituting (A-9) and (A-10) in (A-6), Σ_{MAP} can be calculated according to (A-11):

$$\Pi^{-1} = \frac{1}{2\sigma_x^2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{A-10}$$

$$\Sigma_{MAP} = \frac{1}{2\sigma_x^2+\sigma_i^2+\sigma_j^2}\begin{bmatrix} \sigma_j^2(2\sigma_x^2+\sigma_i^2) & \sigma_i^2\sigma_j^2 \\ \sigma_i^2\sigma_j^2 & \sigma_i^2(2\sigma_x^2+\sigma_j^2) \end{bmatrix} \tag{A-11}$$

Therefore, the relative contribution of prior and likelihood i.e. W_P and W_{MLE} in (A-7), can be drawn as the following equations:

$$W_{MLE} = \frac{1}{2\sigma_x^2+\sigma_i^2+\sigma_j^2}\begin{bmatrix} 2\sigma_x^2+\sigma_j^2 & \sigma_i^2 \\ \sigma_j^2 & 2\sigma_x^2+\sigma_i^2 \end{bmatrix} \tag{A-12}$$

$$W_P = \frac{1}{2\sigma_x^2+\sigma_i^2+\sigma_j^2}\begin{bmatrix} \sigma_i^2 & -\sigma_i^2 \\ -\sigma_j^2 & \sigma_j^2 \end{bmatrix} \tag{A-13}$$

More interestingly, the sum of W_{MLE} and W_P is equal to identity matrix, i.e. $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. This linear combination of S_p and $\hat{\mathrm{S}}^{MLE}$, resembles the way we compute Maximum-A-Posterior estimate by using weighted-averaging of sensory evidence and sensory prior (for more detail see Section 2.3.2).

[54] To compute MAP, an un-normalized form of posterior distribution is enough. So the normalization factor of Byes rule is neglected in the equation.

To analysis the sensitivity of the final estimate to system parameters including: prior variance σ_x^2, and sensory variance (σ_i^2, σ_j^2), we would substitute (A-12) and (A-13) in (A-7), and then, expand its right-hand-side. As a result, the components of $\mathrm{S}^{MAP} = (S_i^{MAP}, S_j^{MAP})$ can be derived as bellow equation:

$$\mathrm{S}^{MAP} = \begin{bmatrix} S_i^{MAP} \\ S_j^{MAP} \end{bmatrix} = \frac{1}{2\sigma_x^2+\sigma_i^2+\sigma_j^2} \left\{ \begin{bmatrix} (2\sigma_x^2 + \sigma_j^2)S_i + \sigma_i^2 S_j \\ \sigma_j^2 S_i + (2\sigma_x^2 + \sigma_i^2)S_j \end{bmatrix} + \begin{bmatrix} \sigma_i^2 (S_i^P - S_j^P) \\ \sigma_j^2 (S_j^P - S_i^P) \end{bmatrix} \right\} \tag{A-14}$$

We assume the priori relation between sensory signals is an identity function, i.e. $S_i^P = S_j^P$. As a result, MAP estimate becomes independent from mean of coupling-prior S^P:

$$\mathrm{S}^{MAP} = \begin{bmatrix} S_i^{MAP} \\ S_j^{MAP} \end{bmatrix} = \frac{1}{2\sigma_x^2+\sigma_i^2+\sigma_j^2} \begin{bmatrix} (2\sigma_x^2 + \sigma_j^2)S_i + \sigma_i^2 S_j \\ \sigma_j^2 S_i + (2\sigma_x^2 + \sigma_i^2)S_j \end{bmatrix} \tag{A-15}$$

Given (A-15), let us analyze the behavior of the system in two extreme cases, where the prior variance σ_x^2 approaches to infinity or zero:

$$\mathrm{S}^{MAP} = \begin{cases} \begin{bmatrix} S_i \\ S_j \end{bmatrix} & \text{if } \sigma_x^2 \to \infty \\ \begin{bmatrix} \frac{S_i\sigma_j^2 + S_j\sigma_i^2}{\sigma_i^2 + \sigma_j^2} \\ \frac{S_i\sigma_j^2 + S_j\sigma_i^2}{\sigma_i^2 + \sigma_j^2} \end{bmatrix} & \text{if } \sigma_x^2 \to 0 \end{cases} \tag{A-16}$$

In case the prior variance approaches to infinity, the MAP estimate becomes identical to MLE. This is dictated by the assumption of no-coupling between signals, and that results in full-segregation. On contrary, when $\sigma_i^2 \to 0$ this implies a certain and bias-free[55] mapping between signals which leads to a full-fusion estimate.

To examine the sensitivity of MAP components to prior variance, the partial derivative of S^{MAP} with respect to σ_x^2 is calculated and shown in equation (A-17):

$$\frac{\partial \mathrm{S}^{MAP}}{\partial \sigma_x^2} = \frac{1}{(2\sigma_x^2+\sigma_i^2+\sigma_j^2)^2} \begin{bmatrix} (S_i - S_j)\sigma_i^2 \\ (S_j - S_i)\sigma_j^2 \end{bmatrix} \tag{A-17}$$

The noticeable fact we can imply from (A-17) is that, the ratio of changing in MAP components, is independent of σ_x^2 and is equal to $\frac{\sigma_i^2}{\sigma_j^2}$. On the other hand, the sensitivity corresponds to each component, is proportional to sensory discrepancy i.e. $\mathrm{D}^{MLE} = (S_i - S_j)$. Greater the sensory discrepancy becomes, faster the MAP estimate changes for both components.

[55] The width of prior variance reflects the probability of possible sensory-discrepancies, both bias-driven and noise-driven.

Appendix B

A practical case study for designing a discrete Extended Kalman Filter

B.1 Problem Definition: Tracking a moving subsurface target using Sonar sensor and EKF

In Section 2.3.4 we introduced the dynamic Bayesian Models of Sensor Fusion. Kalman Filter is the most well-known type of these models in which the dynamics of a system can be estimated in time. Given a sensory evidence (z_k) of the hidden state variables, and the previous state of the system x_k, KF can optimally estimate the state variables at next time step x_{k+1} This process is the first step of the KF algorithm and is called Prediction or State

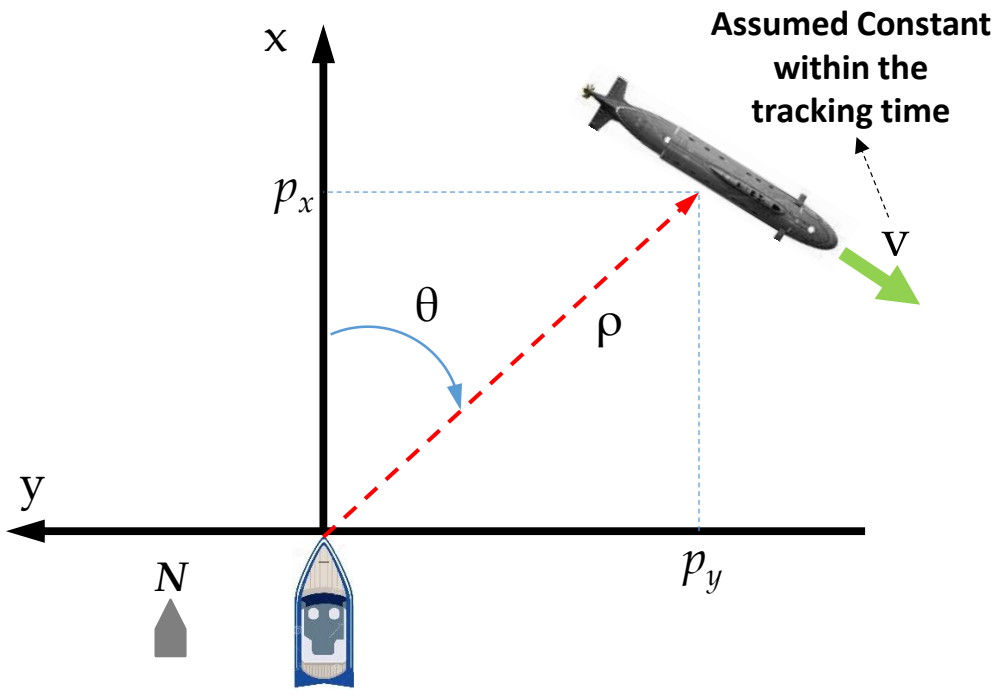


FIGURE B-1
The scheme of the problem setup, including a boat equipped with a Sonar sensor to detect and track subsurface targets.

Transition. After state prediction, the sensory information z_k, will be used to update and compensate the predicted state. This phase is the second step of KF algorithm and is called Measurement Update. Thereby two mapping functions either linear or non-linear should be defined to model the process of State Transition and Measurement Update. For example, in FIGURE 2-11, to estimate x_{k+1} from x_k and u_{k+1}, mapping functions F and B (can be also conditional probability functions) are defined. The state variables are not directly observable, and we have only a noisy sensory evidence associated with each state. To check whether the estimated state is compatible with the sensory evidences, the second mapping H (can also be seen as a conditional probability function), is defined to predict the sensory signal which is likely derived at the estimated state vector x_{k+1}. Then, by comparing the predicted sensory signals with real sensory observation, we are able to compensate the estimated state x_{k+1}. KF is an iterative process, by which the outputs of the previous iteration are the inputs to the next (FIGURE 2-11). This style of information fusion allows the filter to converge towards a more accurate estimate and to cancel out the perturbations caused by intrinsic noise or systematic bias. In case the mapping functions are nonlinear, the KF algorithm is referred as Extended Kalman Filter in which the nonlinear functions are usually linearized by using Tylor expansion around the current state or current sensory observation.

The question of how to derive and identify the parameters of an EKF given a problem, is about to be addressed and answered through this Appendix. We have defined a hypothetical and practical case study to demonstrate how to design an Extended Kalman Filter and to identify its parameters for the proposed problem. The problem consists a boat which is equipped with a noisy sonar sensor. The sonar provides two signals about subsurface targets: the range ρ, and the angel of azimuth θ. In FIGURE B-1 we can see the setup of the problem where a boat is heading north and a submarine is moving under the surface with constant velocity V, and we are going to track the target using sonar sensor which is subject to a heavy white noise, and to cancel out the white noise, and to ultimately track the trajectory of the target using EKF algorithm. There are two assumptions in this problem:

- The sonar sensor fluctuation is modeled by a white Gaussian noise process.
- The norm of velocity vector is constant over time of experiment.

B.2 Formulating the Filter Parameters

B.2.1 Prediction Phase (State Transition)

State vector in kth time step is defined as $X_k = \begin{bmatrix} x_k \\ y_k \\ v_x \\ v_y \end{bmatrix}$ where the x_k and y_k are position of the target in Cartesian coordinate, and v_x and v_y are the components of constant velocity in Cartesian coordinate. Since the velocity is assumed to be constant, and there is no

control input u_k in the system, so we can define the Prediction Equations as following equations (B.1) and (B.2):

$$X_k = \begin{bmatrix} x_{k-1} + v_x T_s \\ y_{k-1} + v_y T_s \\ v_x \\ v_y \end{bmatrix} = \begin{bmatrix} 1 & 0 & T_s & 0 \\ 0 & 1 & 0 & T_s \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{k-1} \\ y_{k-1} \\ v_x \\ v_y \end{bmatrix} \rightarrow X_k^p = F\tilde{X}_{k-1}; F = \begin{bmatrix} 1 & 0 & T_s & 0 \\ 0 & 1 & 0 & T_s \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{(B.1)}$$

Where $\tilde{X}_{k-1}$ the estimated state vector at previous time step, and F is State Transition Matrix, Where T_s is sampling time and set to 0.1 sec.

To measure the quality of prediction estimate, it is necessary to estimate the covariance matrix of the state variables. This matrix describes the uncertainty of the state estimate at each time step, and the correlation between each state variable. In prediction phase, we can give an initial estimate of this matrix P_k^p, that is also defined according to equation (B.2). Note that this estimation will be updated in next phase and is not the final estimate:

$$P_k^p = F\tilde{P}_{k-1}F^T + Q \quad \text{(B.2)}$$

$$Q = q \begin{bmatrix} T_s^3/3 & 0 & T_s^2/2 & 0 \\ 0 & T_s^3/3 & 0 & T_s^2/2 \\ T_s^2/2 & 0 & T_s & 0 \\ 0 & T_s^2/2 & 0 & T_s \end{bmatrix} \quad \text{(B.3)}$$

The state noise V_{k+1} is a random vector that represents the process noise with Normal distribution. Q in equation (B.3) is the covariance matrix of V_{k+1}, and q is the positive scaling factor that indicates the strength of the process noise.

Since EKF is an iterative process, P_{k+1} must be initialized. Initialization is an important task because that will affect the behavior and the convergence of the filter. The diagonal elements of P must represent the variances of each associated state vector element. In case the initial state is at hand, P can be usually initialized to all zeros. This allows EKF to use the initial state estimate to compensate initial noisy sensor observations. However, if the initial state is not known, the diagonal elements should be set to a higher value so that the initial state values do not influence the estimate significantly. Since we have initial state at hand, we have initialized P by zero Matrix; $P_0 = 0$.

B.2.2 Sensory Measurement-Update

In this phase EKF combines the sensory observation, with the information provided by prediction phase, to reduce the uncertainty and thereby to give an optimal state estimation. The factor of prediction phase contribution, and sensory observation is defined by Kalman Gain, K which is computed equation (B.4):

$$K_k = P_k^p H^T \left(HP_k^p H^T + R\right)^{-1} \quad \text{(B.4)}$$

In equation (B.3), we already have calculated P_k^p according to equation (B.2). The **observation Matrix**, $\boldsymbol{H}$ is the mapping matrix that relates the state vector X_k to the sensory observation Z_k; (see FIGURE 2-11). If this mapping is nonlinear, e.g. $h: R^n \rightarrow R^m, Z_k =$

$h(X_k, w_k)$, using Taylor expansion and linearization around current state X_{k-1}, Matrix H can be derived. In this case H is the Jacobian Matrix of $h(X_k, w_k)$ for estimated $\tilde{X}_k$ in equation (B.1). w_k models the noise process governing sensory observation (see equation (B.5)). In equation (B.4), R is sensory noise covariance matrix. In current problem, the observation mapping is nonlinear, because the sonar sensor is not directly measuring the momentary position of the target in Cartesian coordinate, and that needs to be transformed from polar coordinate to Cartesian coordinate (*w_k is assumed to an additive Gaussian noise with Covariance Matrix R*):

$$Z_k = \begin{bmatrix}\rho_k \\ \theta_k\end{bmatrix} = \begin{bmatrix}\sqrt{x_k^2 + y_2^2} \\ Arctan(y_k/x_k)\end{bmatrix} + \begin{bmatrix}w_k^\rho \\ w_k^\theta\end{bmatrix} \rightarrow Z_k = h(X_k; w_k) = \begin{bmatrix}h_1(X_k; w_k) \\ h_2(X_k; w_k)\end{bmatrix} \quad \text{(B.5)}$$

Where $w_k = \begin{bmatrix}w_k^\rho \\ w_k^\theta\end{bmatrix}$ is the sensory observation noise vector with covariance matrix $R = \begin{bmatrix}\sigma_\rho^2 & 0 \\ 0 & \sigma_\theta^2\end{bmatrix}$. H_k can be calculated by computing the Jacobian of h as follows:

$$H_k = \begin{bmatrix}x_k/\sqrt{x_k^2 + y_2^2} & y_k/\sqrt{x_k^2 + y_2^2} & 0 & 0 \\ -y_k/x_k^2 + y_2^2 & x_k/x_k^2 + y_2^2 & 0 & 0\end{bmatrix} = \begin{bmatrix}\cos(\theta) & \sin(\theta) & 0 & 0 \\ -\sin(\theta)/\rho & \cos(\theta)/\rho & 0 & 0\end{bmatrix} \quad \text{(B.6)}$$

Moreover, since the observation noise vector $w_k \sim N(0, R)$ is also transformed through a nonlinear function, we should compute the covariance Matrix of transformed variables by the following equation:

$$R' = W_k R W_k^T;\ W_k = \frac{\partial h(0; w_k)}{\partial w_k} = \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix} \rightarrow R' = R \quad \text{(B.7)}$$

Since W_k is an identical matrix in this problem, consequently $R' = R$. Having EKF gain K_k calculated, now we can update and compensate the state vector and associated covariance matrix P according to equations (B.8) and (B.9):

$$\tilde{X}_k = X_k^p + K_k(Z_k - h(X_k^p; 0)) \quad \text{(B.8)}$$

$$\tilde{P}_k = (I - K_k H_k) P_k^p \quad \text{(B.9)}$$

To sum up, the equation (B.1) and (B.2) are EKF prediction phase equations, and equations (B.4), (B.8) and (B.9) are measurement update phase equations. We have also derived intermediate matrices including state transition matrix and observation matrix for the problem.

B.3 Simulation and Analysis:

In this part of the case study, we have developed a piece of code in MATLAB, to generate 100 hypothetical data set of the Sonar read out, through the trajectory of the

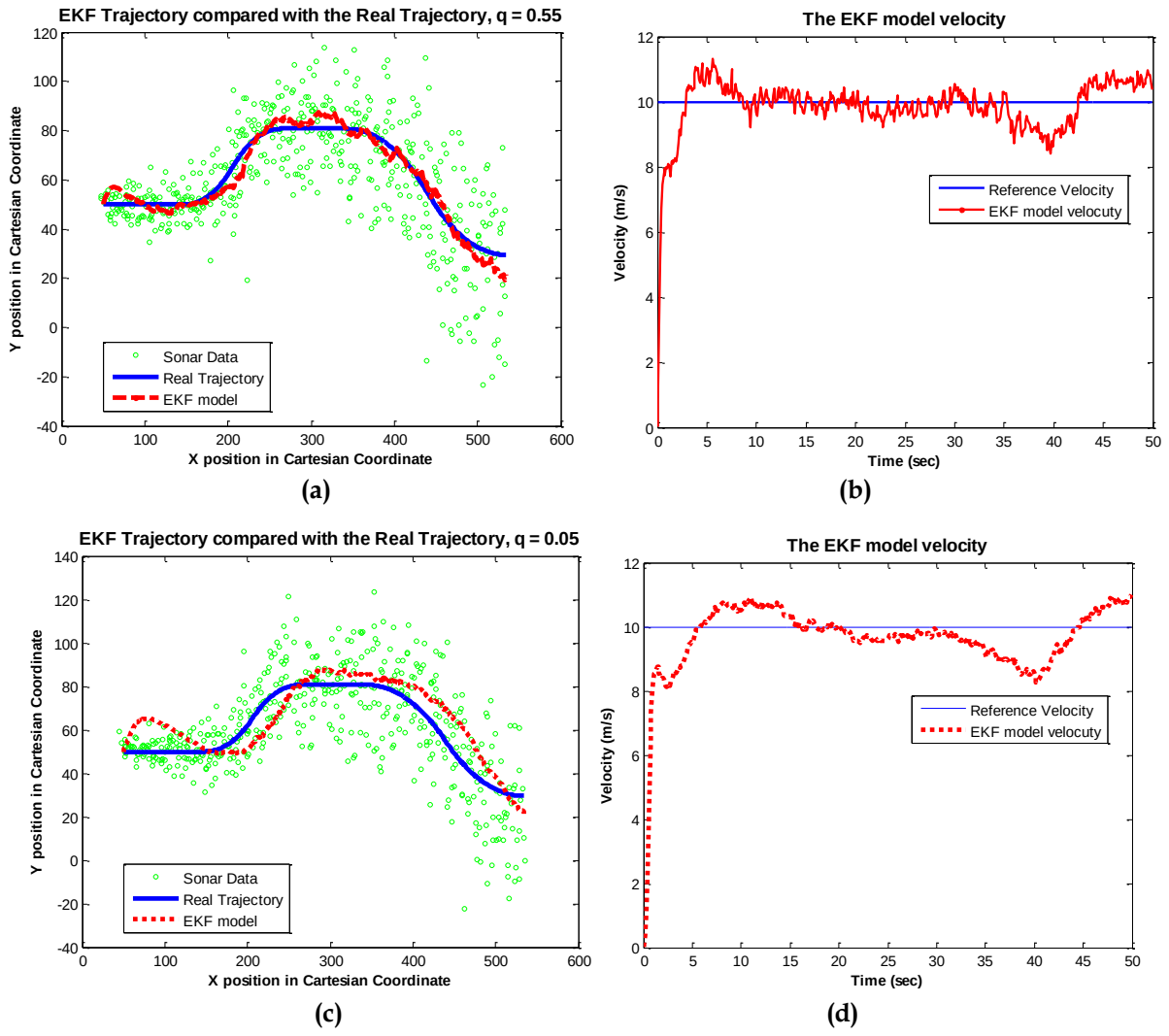


FIGURE B-2

(a) and (c): the real trajectory of submarine is compared with and EKF model. (b) and (d): the reference velocity compared with EKF model velocity. The value of q in top figures is set to 0.55, and in bottom figures it is set to 0.05.

Submarine (`SonarDataSetGen.m`)[56]. This function is using equation (B.5) to transform the real location of the target from Cartesian coordinate which is at disposal, to polar coordinate. Then, we have added identically independent vectors of noise to each point of the trajectory (w_k in equation (B.5) $w_k \sim N(0,R)$). Each of these 100 data set is the input to the EKF algorithm to estimate the state transition of the target X_k.

B.3.1 EKF implementation:

[56] All MATLAB scripts and figures can be found in https://github.com/AMFtech/EKF_Design_CaseStudy-

We have implemented *EKF* algorithm within *MATLAB* script: **myEKF.m**. In fact this script follows equations (B.1)-(B.4), and (B.6), (B.8), (B.9) and we have initialized the matrices including F, Q and R according to system properties and the mentioned assumptions. As we discussed in first section, we set $P_0 = 0$. The initial velocity components are also assumed to be zero. In FIGURE B.2 the behavior of the filter in response to a hypothetical sensory observation (green dot in FIGURE B.2–(a)) is illustrated. Note that the sensory read-out for a single sample trajectory is extremely noisy. However as we can see in FIGURE B.2-(a), the filter is able to track the state of the target, and to reduce the fluctuation drastically. FIGURE B.2-(b) shows the velocity of the target detected by the filter which is still fluctuating around reference value, and starts to drift out when the target starts to turn. In top figures the value of q in equation (B.2) that scales the Q matrix, is set to 0.55. If we reduce this value by one order of magnitude (FIGURE B.2-(c), (d)), we have in fact reduced the uncertainty of state vector elements or equivalently the state covariance matrix Q. So as a consequence the Kalman Filter will trust the estimate given by prediction phase (see equation (B.1)) rather than the one that is given by sensory observation. Thereby the final trajectory is less fluctuating, but in cost of a considerable bias (compare FIGURE B.2-(a) and (c) where the EKF trajectory is smoother in case q = 0.05 but it is drastically drifted from reference trajectory). It is worth to mention, to test developed EKF function, it is just enough to set parameters in script **myEKF_TES.m** and run this script to see the performance of EKF for single noisy trajectory.

Appendix C

Reformulating Posterior Probability of Causal Hypothesis in Audio-Visual Perception

Given the Bayesian generative model of FIGURE 5-2, we have derived the equations necessary to compute the posterior probability function of Causal hypothesis, and decision threshold, as a function of multisensory observations. The extracted formula is based on the following assumptions:

1. Stimuli are assume to be uniformly distributed $s \sim U\left[\frac{-a}{2}, \frac{a}{2}\right]$, where $a = \pi$ is the range of stimuli.
2. Sensory noise is additive Gaussian: $x_a \sim N[s, \sigma_a]$ and $x_v \sim N[s, \sigma_v]$.
3. The prior probability of Causal hypothesis is depicted by $P_{com} = P(C = 1)$.
4. The fluctuation of sensory signals x_a and x_v, or equivalently the standard deviation of noise process, is assumed small enough compared with the range of stimuli $\sigma_a, \sigma_v \ll \frac{a}{2}$.

$$P(C = 1|x_a, x_v) = \frac{P(x_a, x_v|C = 1)P_{com}}{P(x_a, x_v|C = 1)\,P_{com} + P(x_a, x_v|C = 2)(1-P_{com})} \tag{C-1}$$

$$P(C = 2|x_a, x_v) = \frac{P(x_a, x_v|C = 2)(1-P_{com})}{P(x_a, x_v|C = 1)\,P_{com} + P(x_a, x_v|C = 2)(1-P_{com})} \tag{C-2}$$

$$\frac{P(C = 1|x_a, x_v)}{P(C = 2|x_a, x_v)} = \frac{P(x_a, x_v|C = 1)}{P(x_a, x_v|C = 2)}\ \frac{Pcom}{1-Pcom} \Rightarrow PR = LR\frac{Pcom}{1-Pcom} \tag{C-3}$$

$$LR = \frac{\int_{-\frac{a}{2}}^{\frac{a}{2}} P(x_a|s)P(x_v|s)P(s|C = 1)ds}{\left[\int_{-\frac{a}{2}}^{\frac{a}{2}} P(x_a|s_a)P(s_v|C = 1)ds_a\right]\left[\int_{-\frac{a}{2}}^{\frac{a}{2}} P(x_v|s_v)P(s_v|C = 1)ds_v\right]} \tag{C-4}$$

$$LR = \frac{\frac{1}{a}\int_{-\frac{a}{2}}^{\frac{a}{2}} P(x_a|s)P(x_v|s)ds}{\frac{1}{a^2}\left[\int_{-\frac{a}{2}}^{\frac{a}{2}} P(x_a|s_a)ds_a\right]\left[\int_{-\frac{a}{2}}^{\frac{a}{2}} P(x_v|s_v)ds_v\right]} \tag{C-5}$$

The denominator term of Likelihood-Ratio is approximately equal to 1, since the integral of a probability distribution function over the range of sensory signal i.e. $\left[-\frac{\pi}{2}, +\frac{\pi}{2}\right]$ is equal to unity. So LR is approximately equal to the numerator of equation (C-5) that can be derived by equations (C-6), (C-7) and (C-8):

$$LR = a \int_{-\frac{a}{2}}^{\frac{a}{2}} P(x_a|s)P(x_v|s)ds = a \int_{-\frac{a}{2}}^{\frac{a}{2}} \frac{1}{\sqrt{2\pi}\sigma_v} e^{\left(\frac{-(s-x_v)^2}{2\sigma_v^2}\right)} \frac{1}{\sqrt{2\pi}\sigma_a} e^{\left(\frac{-(s-x_a)^2}{2\sigma_a^2}\right)} ds \quad \text{(C-6)}$$

By substituting $s - x_v = \tau$, and $x_a - x_v = d$ in right-side of equation (C-6), LR can be reformulated as convolution of two zero-mean Gaussian functions with σ_a and σ_v:

$$LR = a \int_{-\frac{a}{2}}^{\frac{a}{2}} \frac{1}{\sqrt{2\pi}\sigma_v} e^{\left(\frac{-(d-\tau)^2}{2\sigma_v^2}\right)} \frac{1}{\sqrt{2\pi}\sigma_a} e^{\left(\frac{-\tau^2}{2\sigma_a^2}\right)} d\tau = a \left[\frac{1}{\sqrt{2\pi}\sigma_v} e^{\left(\frac{-d^2}{2\sigma_v^2}\right)}\right] * \left[\frac{1}{\sqrt{2\pi}\sigma_a} e^{\left(\frac{-d^2}{2\sigma_a^2}\right)}\right] \quad \text{(C-7)}$$

Convolution of two zero-mean Gaussian functions is equal to another zero-mean Gaussian function whose variance is equal the sum of convolving functions' variances:

$$LR \approx a \left[\frac{1}{\sqrt{2\pi(\sigma_a^2+\sigma_a^2)}} e^{\left(\frac{-d^2}{2(\sigma_a^2+\sigma_a^2)}\right)}\right] = a\, Q(d) \quad \text{(C-8)}$$

As we can see in (C-8), Likelihood-Ratio is described as a symmetric function of spatial disparity between Audio-Visual signals. Now substituting (C-8) in (C-3), Posterior-Ratio can be calculated:

$$PR = \frac{P(C=1|x_a,x_v)}{P(C=2|x_a,x_v)} = \text{a}\, Q(d) \frac{Pcom}{1-Pcom} \quad \text{(C-9)}$$

Similar to *LR, PR* is also a Gaussian function of Audio-Visual disparity, weighted with a homographic term of causal prior P_{com}.

The probability of common-source scenario $P(C=1|x_a,x_v)$, can be specifically derived from Posterior-Ratio, where $d_{av} = |x_a - x_v|$:

$$\frac{P(C=1|x_a,x_v)}{1-P(C=1|x_a,x_v)} = a\, Q(d_{av}) \frac{Pcom}{1-Pcom} \quad \text{(C-10)}$$

$$P(C=1|x_a,x_v) = \frac{a\, Q(d_{av})\, Pcom}{a\, Q(d_{av}) Pcom + 1 - Pcom} \quad \text{(C-11)}$$

If we find the root of Log-PR, that implies the situation at which probability of $P(C=1|x_a,x_v)$ is identical to $P(C=2|x_a,x_v)$, and thereby specifies the decision threshold d_{th}:

$$\log(PR) = \log(\text{a}) + \log\left(\frac{Pcom}{1-Pcom}\right) - \frac{1}{2}\log\left(2\pi(\sigma_a^2+\sigma_a^2)\right) - \frac{1}{2(\sigma_a^2+\sigma_a^2)} d^2 = 0 \quad \text{(C-12)}$$

$$d_{th} = \sqrt{2(\sigma_a^2+\sigma_a^2)\left[\log\left(\frac{a\, Pcom}{1-Pcom}\right) - (0.5)\log\left(2\pi(\sigma_a^2+\sigma_a^2)\right)\right]} \quad \text{(C-13)}$$

References

1. Alias D, Burr D, (2004a), "No direction-specific bimodal facilitation for audiovisual motion detection", Cognitive Brain Research, Vol.19, pp 185–194, 2004
2. Alias D, Burr D, (2004b), "the ventriloquist effect results from near-optimal bimodal integration", Current Biology, Vol.14, pp 257-262, 2004
3. Alvarado J C, Stanford T R, Vaughan J W, Stein B E (2007), "Cortex mediates multisensory but not uni-sensory integration in superior colliculus", Journal of Neuroscience, Vol. 27(47), pp 12775 – 86, November 2007
4. Amari S (1977), "Dynamics of pattern formation in lateral inhibition type neural fields", Biological Cybernetics, Vol.27, pp 77-87, 1977.
5. Avillac M, Deneve S, Olivier E, Pouget A, and Duhamel J R, (2005), "Reference frames for representing visual and tactile locations in parietal cortex". Nature Neuroscience, Vol.8, Issue.7, pp 941–949, 2005
6. Aydin I H, (2001), "Avicenna and modern neurological sciences", Journal of Academic Researches in Religious Sciences, Vol.1, pp 1-4, 2001
7. Axenie C, Conradt J (2013) "Cortically Inspired Sensor Fusion Network for Mobile Robot Heading Estimation", Proc. of International Conf. on Artificial Neural Networks-ICANN13, pp. 240-247, September 2013
8. Axenie C, Conradt J (2015), "Cortically inspired sensor fusion network for mobile robot egomotion estimation", Robotics and Autonomous Systems, Vol.31, pp 69-82, September 2015
9. Balasubramanian S, Melendez-Calderon A, Roby-Brami A, Burdet E (2015), "On the analysis of movement smoothness", Journal of Neuro-Engineering and Rehabilitation, Vol.12 (112), December 2015
10. Banks M S, Burge J, Held R T, (2011), "The Statistical Relationship between Depth, Visual Cues, and Human Perception", In: Sensory Cue Integration, 2nd edition, Oxford University Press, pp 195-223, 2011. Alias D, Burr D, (2003), "The Flash-Lag effect occurs in audition and cross modally", Current Biology, Vol. 13, pp 59-63, Jan 2003
11. Bauer J, Weber C and Wermter S (2012), "A SOM-based model for multi-sensory integration in the superior colliculus", International Joint Conference on Neural Networks 2012, (IJCNN12), pp 1–8, Brisbane, Australia, June 2012
12. Beck J M, Latham P E and Pouget A (2011), "Marginalization in Neural Circuits with Divisive Normalization", Journal of Neuroscience, Vol.31, Issue 43, pp 15310-15319, October 2011
13. Berger D R, (2006), "Sensor Fusion in the Perception of Self-Motion", PhD dissertation, University of Ulm, pp 6-10, 2006
14. Bichler O, Querlioz D, Thorpe S J, Bourgoin J P, and Gamrat C (2012), "Extraction of temporally correlated features from dynamic vision sensors with spike-timing-dependent plasticity" Neural Networks, vol.32 pp 339-348, August 2012.
15. Bisley J W, (2011), "The neural basis of visual attention", Journal of Physiology, Vol. 589, pp 49-57, January 2011

16. Bresciani, J P, Dammeier F, Ernst M O, (2006). "Vision and touch are automatically integrated for the perception of sequences of events", Journal of Vision, Vol.6 Issue 5, pp 554–564, April 2006
17. Bradshaw M F, Rogers B J, (1996), "The interaction of binocular disparity and motion parallax in computation of depth", Vision Research, Vol. 36, pp 3457-3468, 1996
18. Britten K H (2003), "The middle temporal area: motion processing and the link to perception", The Visual Neurosciences, ed. Chalupa L M, Werner JF, pp. 1203–16. MIT Press, Cambridge, MA, USA, Nov 2003
19. Britten K H (2008) "Mechanisms of Self-Motion Perception", Annual Review of Neuroscience. Vol.31, pp 389-410, March 2008
20. Bruce V, Green P R, Mark A (2003), "Visual Perception: Physiology, Psychology, & Ecology" Psychology Press, USA, pp.180-183, (2003)
21. Bruce N D B, Tsotsos J K (2009), "Saliency, Attention, and Visual Search: An Information Theoretic Approach", Journal of Vision, Vol.9 (3), pp 1-24, 2009
22. Boghossian, P, (2014). "What is inference?" Philosophical Studies, Vol 169 No.1, pp 1-18, (2014).
23. Born R T, Bradley D C (2005), "Structure and Function of Visual Area MT", Annual Revision Neuroscience, Vol. 28, pp 157-189, March 2005
24. Boyle S C, Kayser S J, Kayser C, (2017), "Neural correlates of multisensory reliability and perceptual weights emerge at early latencies during audio-visual integration", European Journal of Neuroscience, Vol. 46, Issue 10, pp 2565-2577, November 2017
25. Brostek L, Büttner U, Mustari M J, Glasauer S (2015), "Eye Velocity Gain Fields in MSTd During Optokinetic Stimulation, *Cerebral Cortex*, Vo.25 (8), pp 2181–2190, August 2015
26. Bubic A, Von Cramon D Y, Schubotz R I (2010), "Prediction, Cognition and the Brain", Frontiers in Human Neuroscience, Vol.4 (25), March 2010
27. Burg J, Ernst M O, Banks M S, (2008), "The statistical determinants of adaptation rate in human reaching", Journal of Vision, Vol. 8 (20), pp 1-19, April 2008
28. Calvert G A, Thesen T (2004), "Multisensory integration: methodological approaches and emerging principles in the human brain", Journal of Physiology-Paris, Vol.98, Issues 1–3, pp 191-205, January–June 2004
29. Carandini M, Heeger D J (2012), "Normalization as a canonical neural computation", Nature Reviews Neuroscience, Vol. 13, pp 51- 62, January 2012
30. Carneiro J, Ieng S, Posch C, Benosman R (2013), "Event-based 3D reconstruction from neuromorphic retina", Neural Network, Vol.45, pp 27–38, September 2013
31. Camuñas-Mesa L A, Serrano-Gotarredona T, Ieng S H, Benosman R, Linares-Barranco B (2014), "On the use of orientation filters for 3D reconstruction in event-driven stereo vision", Frontiers Neuroscience, Vol.8 (48), 2014
32. Chen A, DeAngelis G C, Angelaki D E (2013), "Functional Specializations of the ventral intraparietal area for multisensory heading discrimination", Journal of Neuroscience, Vol.33, pp 3567-3581, February 2013

33. Clarke E, O'Malley C D, (1996), "The Human Brain and Spinal Cord: A Historical Study Illustrated by Writings from Antiquity to the Twentieth Century", 2nd edition, Norman Publishing, San Francisco, USA, 1996
34. Cohen Y E, Andersen R A (2002), "A common reference frame for movement plans in the posterior parietal cortex", Nature Reviews Neuroscience, Vol.3, pp 553–562, July 2002
35. Cook M, Jug F, Krautz C, Steger A (2010), "Unsupervised learning of relations", Proc. of International Conference on Artificial Neural Networks – ICANN10, pp. 164-173, Thessaloniki, Greece, September 2010
36. Cook M, Gugelman L, Jug F, Krautz C, Steger A (2011), "Interacting maps for fast visual interpretation", Proc. of International Joint Conference on Neural Networks – IJCNN11, pp. 770-776, San Jose, California, USA, August 2011
37. Cooper R P, Shallice T (2006), "Hierarchical schemas and goals in the control of sequential behavior", Psychological Review, Vol.113, pp 887–916, 2006
38. Conradt J, Pescatore M, Pascal S, Verschure P F M J (2002), "Saliency Maps Operating on Stereo Images Detect Landmarks and their Distance", International Conference on Artificial Neural Networks (ICANN), Madrid, Spain, 2002, pp.795-800, (2002)
39. Conradt J, Cook M, Berner R, Lichtsteiner P, Douglas R J, Delbruck T (2009), "A pencil balancing robot using a pair of AER dynamic vision sensors", IEEE International Symposium on Circuits and Systems, Taipeh, Taiwan, pp 781-784, May 2009
40. Conradt J, Galluppi F, Stewart T C (2015), "Trainable sensorimotor mapping in a neuromorphic robot", Robotics and Autonomous Systems, Vol.71, pp 60-68, September 2015
41. Conway C M, and Christiansen M H, (2006), "Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations", Psychological Science, Vol. 17, pp 905–912, October 2006.
42. Cruse H, Wischmeyer E, Brüwer M, Brockfeld P, Dress A (1990),"On the cost functions for the control of the human arm movement", Biological Cybernetics, Vol.62 (6), pp 519–528, April 1990
43. Das A (2005), "Cortical Maps: Where Theory Meets Experiments", Vol. 47 (2), pp 168-171, July 2005
44. Desovski D, Liu Y, Cukic B (2005), "Linear randomized voting algorithm for fault tolerant sensor fusion and the corresponding reliability model", Ninth IEEE International Symposium on High-Assurance Systems Engineering (HASE'05), Heidelberg, Germany, October 2005
45. Dikov G, Firouzi M, Röhrbein F, Conradt J, Richter C (2017), "Spiking Cooperative Stereo-Matching at 2 ms Latency with Neuromorphic Hardware", In: Mangan M., Cutkosky M., Mura A., Verschure P., Prescott T., Lepora N. (eds) Biomimetic and Biohybrid Systems. Living Machines 2017, Vol 10384, Springer, pp 119-137, July 2017
46. Drazen D, Lichtsteiner P, Hafliger P, Delbruck T, Jensen A (2011), "Toward real-time particle tracking using an event-based dynamic vision sensor", Experiments in Fluids, Vol. 51, pp 1465-1469, 2011

47. Elman J L, (1990). "Finding Structure in Time". Cognitive Science, Vol.14, Issue 2, pp 179–211, 1990
48. Ernst M O, Banks M S, (2002), "Humans integrate visual and haptic information in a statistically optimal fashion", Nature, Vol.415, pp 429-433, 2002
49. Ernst M. O., Bülthoff H. H., (2004), "Merging the senses into a robust percept", Trends on Cognitive Science, Vol.8, pp 162-168, April 2004.
50. Ernst, M O, (2005). "A Bayesian view on multimodal cue integration" (chapter 6). In G. Knoblich, M. Grosjean, I. Thornton, & M. Shiffrar (Eds.), Human body perception from the inside out (pp. 105–131). New York, USA: Oxford University Press.
51. Ernst M O, (2007). "Learning to integrate arbitrary signals from vision and touch", (2007), Journal of Vision, Vol. 7(7), pp 1-14, June 2007
52. Ernst M O, Di Luca M, (2011), "Multisensory Perception: From Integration to Remapping", in Sensory Cue Integration, edited by Trommershauser, Körding, Landy, Oxford University Press, pp 224-250, 2011
53. Felleman D J, Van Essen D C (1991), "Distributed Hierarchical Processing in the Primate Cerebral Cortex", Cerebral Cortex, Vol.1 pp, 1-47, Feb 1991
54. Fetsch C R, Turner A H, DeAngelis G C, Angelaki D E (2009), "Dynamic Reweighting of Visual and Vestibular Cues during Self-Motion Perception", The Journal of Neuroscience, Vol.29 (49), pp 15601-15612, December 2009
55. Fetsch C R, Pouget A, DeAngelis G C, and Angelaki D E, (2011), "Neural correlates of reliability-based cue weighting during multisensory integration", Nature Neuroscience, Vol. 15, Issue 1, pp 146–154, November 2011
56. Filipe S, Alexandre L A (2015), "From the human visual system to the computational models of visual attention: a survey", Artificial Intelligence Review, Vol. 43 (4), pp 601–601, April 2015
57. Firouzi M, Glasauer S, Conradt J, (2014a), "Flexible Cue Integration by Line Attraction Dynamics and Divisive Normalization", Proc. Of 24th International Conf. of Artificial Neural Network, pp 691-698, Hamburg, Germany, September 2014
58. Firouzi M, Axenie C, Glasauer S, Conradt J (2014b), "Modulating Neural Activity can bias Neural Dynamics in Attractor Networks for Optimal Weighted Cueing", BCCN-Munich Conference, Tutzing, July 2014
59. Firouzi M, Shouraki S B, Afrakoti I E P (2014c), "Pattern analysis by active learning method classifier", Journal of Intelligent & Fuzzy Systems, Vol. 26, No. 1, pp 49-62, 2014
60. Firouzi M, Shouraki S B, Conradt J, (2014d), "Sensorimotor Control Learning Using a New Adaptive Spiking Neuro-Fuzzy Machine, Spike-IDS and STDP", Proc. Of 24th International Conf. of Artificial Neural Network, pp 379-386, Hamburg, Germany, September 2014
61. Firouzi M, Glasauer S, Conradt J, (2015) "Causal Bayesian Inference in hierarchical distributed computation in the Cortex, towards a neural model", Bernstein Conference on Computational Neuroscience 2015, Heidelberg, Germany September 2015, DOI: 10.12751/nncn.bc2015.0182

62. Firouzi M, Glasauer S, Conradt J (2016). "Probabilistic Causal Inference in Multisensory Perception, a distributed hierarchical model", Bernstein Conference on Computational Neuroscience 2016, Berlin, September 2016, Doi: 10.12751/nncn.bc2016.0037
63. Firouzi M, Conradt J (2016), "Asynchronous Event-based Cooperative Stereo Matching Using Neuromorphic Silicon Retinas", Neural Processing Letters, 1. Vol.43 (2), pp 311–326, April 2016
64. Fischer B J (2010), "Bayesian Estimation from Heterogeneous Population Codes", The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, July 2010
65. Foucault M, (2004), "History of Madness", English Translation, 1st Edition, Taylor & Francis, London, UK, October 2004
66. Friston K, (2005), "A theory of cortical responses", Philosophical Transactions of The Royal Society B, Biological Sciences, Vol. 360, pp 815-36, May 2005
67. Furber S B, Brown A D (2009), "Biologically-Inspired Massively-Parallel Architectures-computing beyond a million processors", 9th International Conference on Application of Concurrency to System Design (ACSD), USA, pp 3-12, July 2009
68. Galluppi F, Denk C, Meiner M C, Stewart T, Plana L, Eliasmith C, Furber C, Conradt J (2014), "Event-based neural computing on an autonomous mobile platform", IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, pp. 2862-2867, June 2014
69. Ganguli D, Simoncelli E P (2014), "Efficient sensory encoding and Bayesian inference with heterogeneous neural populations" Neural Computation, Vol.26 (10), pp 2103-2134, October 2014
70. Ghahramani Z, Wolptrt D M, Jordan M I, (1997) "Computational models of sensorimotor integration", Advances in Psychology Vol. 119, pp 117-147, 1997
71. Ghahramani Z, (2001), "Introduction to Hidden Markov Models and Bayesian Networks", Journal of Pattern Recognition and Artificial Intelligence, Vol. 15, No. 1, pp 9-42, 2001
72. Girshick A R, Landy M S, Simoncelli E P, (2011), "Cardinal rules: visual orientation perception reflects knowledge of environmental statistics", Nature Neuroscience, Vol. 14, pp 926-32, June 2011
73. Grewel M S, Andrews A P, (2010), "Applications of Kalman Filtering in Aerospace 1960 to the Present" IEEE Control Systems Magazine Vol. 30, Issue 3, June 2010.
74. Grieve K L, Sillito A M, (1995), "Differential properties of cells in the feline primary visual cortex providing the cortico-fugal feedback to the lateral geniculate nucleus and visual claustrum", Journal of Neuroscience, Vol. 15, pp. 4868-4874, 1995
75. Grover Brown R, Hwang Y C P, (2012) "Introduction to Random Signals and Applied Kalman Filtering with MATLAB Exercises", John Wiley & Sons, USA, 4th edition, Feb 2012
76. Gruters K G, Groh J M, (2012), "Sounds and beyond: multisensory and other non-auditory signals in the inferior colliculus", Frontiers in Neural Circuits, December 2012

77. Guo M P, Takezawa T, Allig C, and Firouzi M (2015), "Developing a Recurrent Neural Network solution for motion estimation to enhance attention-based tracking network, Project Lab Report, ECE Dept., Technical University of Munich, July 2015
78. Gustafsson F, (2010), "Particle filter theory and practice with positioning applications", IEEE Aerospace and Electronic Systems Magazine, Vol. 25, Issue. 7, pp 53-82, July 2010
79. Hall D L, Llinas J (2009), "Multisensory Data Fusion", Handbook of Multisensor Data Fusion, Theory and Practice, 2nd Edition, Edited by: Liggins M E, Hall D L, Llinas J, CRC Press, New York, 2009
80. Hartline P H (1988), "Multisensory Convergence", In: Sensory Systems II, Senses Other than Vision - Readings from the Encyclopedia of Neuroscience, pp 47-50, Birkhäuser, Boston, 1988
81. Hartly R, Zisserman A (2003), "Multiple View Geometry in Computer Vision", 2d edition, Cambridge University press, UK, pp 239-261, 2003
82. Haxby J V, Connolly A C, Guntupalli J S (2014), "Decoding neural representational spaces using multivariate pattern analysis", Annual Review Neuroscience, Vol.37, pp 435-456, June 2014
83. Hess S, "Low-level stereo matching using event-based Silicon Retina", Semester project report, ETHZ Zurich, pp. 16, 2006
84. Hoffmann R, Weikersdorfer D, Conradt J (2013), "Autonomous Indoor Exploration with an Event-Based Visual SLAM System" European Conference on Mobile Robots, Barcelona, Spain, pp 38-43, September 2013
85. Hogendoorn H, Burkitt A N, (2018), "Predictive coding of visual object position ahead of motion", Neuroimage , Vol. 171, pp 55-61, May 2018
86. Hosoya T, Baccus S A, Meister M, (2005), "Dynamic predictive coding by the retina", Nature, Vol. 436, pp 71–77, July 2005
87. Humphreys G F, Lambon Ralph M A (2015), "Fusion and Fission of Cognitive Functions in the Human Parietal Cortex", Cerebral Cortex, Vol.25 (10), pp 3547–3560, October 2015
88. Itti, L, Koch C (2001), "Computational modeling of visual attention", Nature Reviews Neuroscience vol.2, pp 1-11, March 2001
89. Indiveri G, Douglas R (2000), "Neuromorphic vision sensors" Science, vol.288, no.5469, pp. 1189–1190, 2000
90. Jazayeri M, Movshon A, "Optimal representation of sensory information by neural populations", Nature Neuroscience Vol.9, pp. 690 – 696, 2006
91. Jórdeczka M, (2016), "Trepanation of the skull before millennia ago in Sudan", Science and Scholarship in Poland, June 2016
92. Jun Z (1991), "Dynamics and formation of self-organizing maps", Neural Computing, Vol. 3(1), pp 54-66, Spring 1991
93. Kayser C, Petkov C I, Logothetis N K (2009), "Multisensory interactions in primate auditory cortex: fMRI and electrophysiology", Hearing Research, Vol.258, pp 80–88, December 2009

94. Kayser C, Shams L (2015), "Multisensory Causal Inference in the Brain", PLOS Biology, Vol.13 (2), February 2015
95. Keil J, Senkowski D (2018), "Neural Oscillations Orchestrate Multisensory Processing", Neuroscientist. Vol.24 (6), pp 609-626, February 2018
96. Kiefer A B, "Literal Perceptual Inference", (2017), Eds. Metzinger & Wiese, Philosophy and Predictive Processing, Chapter 17, Johannes Gutenberg-Universität of Mainz, Frankfurt am main, 2017
97. Knill D C, (2007), "Bobust Cue Integration: A Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant", Journal of Vision, 7(5), pp 1-24, 2007.
98. Körding K P, Wolpert D M, (2006), "Bayesian decision theory in sensorimotor control", Trends in Cognitive Sciences Vol. 10 No. 7, pp 320-326, Jun 2006.
99. Körding K. P, Beierholm U, Ma W. J, Quartz S, Tenenbaum J. B, Shams L, (2007) "Causal Inference in Multisensory Perception", PLOS one, Issue 9, September 2007
100. Kogler J, Humenberger M, Sulzbachner C (2009), "Event-Based Stereo Matching Approaches for Frameless Address Event Stereo Data", Advances in visual Computing, Lecture notes in Computer Science, Vol.6938, pp 674-685, 2011
101. Koyuncu A B, "Evaluation and validation of a motion-cued attention network using a Robot-Head platform" (2016), B.Sc. Dissertation, ECE Dept., Technical University of Munich, September 2016.
102. Krekelberg B, van Wezel R J A, and Albright T D (2006), "Interactions between Speed and Contrast Tuning in the Middle Temporal Area: Implications for the Neural Code for Speed", Journal of Neuroscience Vol.26 (35), pp 8988 - 8998, August 2006
103. Kersten D, Mamassian P, and Yuille A L (2004), "Object perception as Bayesian inference", Annual Review of Psychology, Vol. 55, pp 271–304, 2004
104. Landy M S, Banks M S, Knill D C, (2011), "Ideal-Observer Models of Cue Integration", Sensory Cue Integration, 2nd edition, Oxford University Press, pp 5-29, 2011
105. Lee T S, Yuille A L, (2007), "Efficient Coding of Visual Scenes by regrouping and Segmentation", in Bayesian Brain, a probabilistic approach to neural coding, edited by Doya K, Ishi S, Pouget A, Rao R P N, MIT press, Cambridge, MA, USA, 2007
106. Liang M, Mouraux A, Iannetti G D (2013), "Bypassing primary sensory cortices - a direct thalamocortical pathway for transmitting salient sensory information", Cerebral Cortex, Vol.23, Issue 1, pp 1–11, January 2013
107. Lichtsteiner L, Posch C, Delbruck T A (2008), "A 128×128 120dB 15us Latency Asynchronous Temporal Contrast Vision Sensor", IEEE Journal of Solid State Circuits, Vol.43 (2) pp 566-576, 2008
108. Lindner A , Thier P , Kircher TT , Haarmeier T , Leube DT (2005), "Disorders of Agency in Schizophrenia Correlate with an Inability to Compensate for the Sensory Consequences of Actions", Current Biology, Vol.15 (12), pp 1119-1124, Jun 2005
109. Liu T, Hou Y, (2013), "A Hierarchy of Attentional Priority Signals in Human Frontoparietal Cortex", The Journal of Neuroscience, Vol.33 (42), pp 16606-16616, October 2013

110. Ma W J, Beck J M, Latham P E and Pouget A (2006), "Bayesian inference with probabilistic population codes", Nature Neuroscience, Vol.9, pp 1432–1438, October 2006
111. Ma W J, Rahmati M (2013)," Towards a neural implementation of causal inference in cue combination", Multisensory Research, Vol.26 (1-2), pp 159-176, January 2013
112. Ma W J, Zhou X, Ross, L A, Foxe J, and Parra L C, (2009), "Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space", PLoS One, March 2009
113. Macaluso E (2012), "Spatial Constraints in Multisensory Attention", In: The Neural Bases of Multisensory Processes, Murray M M, Wallace M T, editors, CRC Press/Taylor & Francis, Chapter 25, 2012
114. Magosso E, Cuppini C, Serino A, Di Pellegrino G, Ursino M (2008), "A theoretical study of multisensory integration in the superior colliculus by a neural network model", Neural Network, Vol. 21(6), pp 817-29, June 2008
115. Mahowald M A, Mead C (1991), "The silicon retina", Scientific American 264(5), pp 76-82, May 1991
116. Mahowald M, Delbrück T, "Cooperative stereo matching using static and dynamic image features". In C. Mead and M. Ismail (eds.) Analog VLSI Implementation of Neural Systems, Kluwer Academic Publishers, Boston, USA, pp 213-238, 1989
117. Makin T R, Holmes N P, Zohary E (2007), "Is that near my hand? Multisensory representation of peripersonal space in human intraparietal sulcus", Journal of Neuroscience., Vol.27, pp 731–740, December 2007
118. Marr D, "Vision, A computational Investigation into the Human Representation and Processing of Visual Information" edited by Lucia M. Vaina, MIT press, 2010, pp. 111-122, 2010
119. Martin J G, Meredith M A, Ahmad Kh (2009), "Modeling Multisensory Enhancement with Self-organizing Maps", Frontier in Computational Neuroscience, Vol.3 (8), June 2009
120. Meredith M A, Wallace M T, Stein B E (1992), "Visual, auditory and somatosensory convergence in output neurons of the cat superior colliculus: Multisensory properties of the tecto-reticulo-spinal projection", Experimental Brain Research, Vol. 88, pp 181–186, 1992
121. Mesulam M M (1998), "From sensation to cognition" Brain, Vol. 121, Issue 6, pp 1013–1052, June 1998
122. Miller K D (2016), "Canonical computations of cerebral cortex", (2016), Current Opinions in Neurobiology, Vol. 37, pp 75-84, April 2016
123. Müller G R, Conradt J (2012), "Self-calibrating Marker Tracking in 3D with Event-Based Vision Sensors", 22nd International Conference on Artificial Neural Networks (ICANN), pp 313-321, 2012
124. Musacchia G, Sams M, Nicol T, Kraus N, (2006) "Seeing speech affects acoustic information processing in the human brainstem", Experimental Brain Research., Vol.168 (1-2), pp 1–10, January 2006

125. Myung I J, (2003), "Tutorial on maximum likelihood estimation", Journal of Mathematical Psychology, Vol.47, pp 90-100, 2003
126. Ni Z, Pacoret C, Benosman R, Ieng S H, Régnier S (2012), "Asynchronous event-based high speed vision for microparticle tracking". Journal of microscopy. Vol.245 (3), pp 236–244, 2012
127. Olcese U, Oude-Lohuis M N, and Pennartz C M A (2018), "Sensory Processing Across Conscious and Nonconscious Brain States: From Single Neurons to Distributed Networks for Inferential Representation" Frontiers in Systems Neuroscience, Vol.12 (49), October 2018
128. Osswald M, Ieng S H, Benosman R, Indiveri G (2017), "A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems", Scientific Reports, vol.7, Article 40703, March 2017
129. Paraskevopoulos E, Herholz S C (2013), "Multisensory integration and neuroplasticity in the human cerebral cortex", Translational Neuroscience, Vol.4 (3), pp 337-348, September 2013
130. Parhami B (1996), "A taxonomy of voting schemes for data fusion and dependable computation", Reliability Engineering & System Safety, Vol.52, Issue.2, pp 139-151, May 1996
131. Petersen S E, Posner M I (2012), "The Attention System of the Human Brain: 20 Years After", Annual Review of Neuroscience, Vol.35, pp 73-89, July 2012
132. Petzschner F H, Glasauer S (2011), "Iterative Bayesian Estimation as an Explanation for Range and Regression Effects: A Study on Human Path Integration" Journal of Neuroscience, Vol.31 (47), pp 17220-17229, November 2011
133. Pham D T and Liu X (1996), "Training of Elman networks and dynamic system modelling", International Journal of Systems Science, Vol.27, No.2, pp 221-226, 1996
134. Pitkow X, Liu S, Angelaki D E, DeAngelis G C, and Pouget A (2015), "How Can Single Sensory Neuron Predict Behavior?", Neuron, Vol.87, pp 411-423, July 2015
135. Pouget A, Sejnowski T J (1997), "Spatial Transformations in the Parietal Cortex Using Basis Functions", Journal of Cognitive Neuroscience, Vol.9 (2), pp 222 – 237, March 1997
136. Pouget A, Deneve S, Sejnowski T J (1999), "Frames of reference in hemineglect: a computational approach", Progress in Brain Research, vol.121, pp 81-97, January 1999
137. Pouget A, Snyder L H (2000), "Computational Approaches to sensorimotor transformation", Nature Neuroscience, Vol.3, pp 1192-1198, November 2000
138. Pouget A, Beck J M, Ma W J, Latham P E, (2013), "Probabilistic brains: knowns and unknowns", Nature Neuroscience, Vol. 16, No. 9, pp 1170–1178, Sept 2013
139. Quiroga R Q, Reddy L, Kreiman G, Koch C, Fried I (2005), "Invariant Visual Representation by Single Neurons in the Human Brain", Nature. Vol. 23:435(7045), pp 1102-1107, June 2005
140. Recanzone G H, (2004), "Auditory Influences on Visual Temporal Rate Perception", Journal of Neurophysiology, Vol. 89, pp 1078-1093, 2004
141. Remme M W, Wadman W J (2012), "Homeostatic scaling of excitability in recurrent neural networks" PLoS Computational Biology, Vol.8 (5), May 2012

142. Reynolds J H, Heeger D J, "The normalization model of attention", Neuron, Vol.29, No.61(2), pp 168-185, Jan 2009
143. Riesenhuber M, Paggio T (1999), "Hierarchical models of object recognition in cortex", nature neuroscience, vol.2, No.11, pp 1019-1025, November 1999
144. Roach N, Heron J, McGraw P, (2006), "Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration", Proceedings of the Royal Society of London B: Biological Science, Vol. 273, pp 2159-2168, Sept 2006
145. Rogister P, Benosman R, Ieng S H, Lichtsteiner P, Delbruck T (2012), "Asynchronous Event-Based Binocular Stereo Matching" Neural Networks and Learning Systems, IEEE Transactions on, Vol.23 (2), pp 347–353, 2012
146. Rohe T, Noppeney U (2015), "Cortical Hierarchies Perform Bayesian Causal Inference in Multisensory Perception", PLOS Biology, Vol.13 (2), February 2015
147. Rohe T, Noppeney U, (2016), "Distinct Computational Principles Govern Multisensory Integration in Primary Sensory and Association Cortices", Vol.26, Issue 4, pp 509–514, February 2016
148. Rohe T, Noppeney U (2018), "Reliability-Weighted Integration of Audiovisual Signals Can Be Modulated by Top-down Attention", eNeuron, Vol.5 (1), February 2018
149. Rolls E T, Deco G (2010), "The Noisy Brain Stochastic Dynamics as a Principle of Brain Function", 1st Edition, ISBN 978-0-19-958786-5, Oxford University Press, UK, March 2010
150. Rougier N P (2006), "Dynamic Neural Field with Local Inhibition", Biological Cybernetics, Vol.94 (3), pp 169-179, February 2006
151. Rougier N P, Vitay J, (2006), "Emergence of attention within a neural population", Neural Networks, Vol.19, No.5, pp 573-581, January 2006.
152. Rougier N P, Vitay J, (2011), "Synchronous and asynchronous evaluation of dynamic neural fields", Journal of Difference Equations and Applications, Vol.17, pp 1119-1133, January 2011
153. Saalmann Y B, Pigarev I N, Vidyasagar T R (2007), "Neural mechanisms of visual attention: how top-down feedback highlights relevant locations", Science, Vol.15 (316), pp 1612-15, June 2007
154. Sagha H, Millán J D R, Chavarriaga R (2011), "Detecting anomalies to improve classification performance in opportunistic sensor networks" In 2011 IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 154-159, March 2011
155. Sagha H, Bayati H, Millán J D R, Chavarriaga R (2013), "On-line anomaly detection and resilience in classifier ensembles", Pattern Recognition Letters, Vil.34 (15), pp 1916-1927, 2013
156. Sandamirskaya Y (2014), "Dynamic neural fields as a step toward cognitive neuromorphic architectures", Frontiers in Neuroscience, Vol.23, January 2014
157. Sato Y, Toyoizumi T, Aihara K (2007), "Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli", Neural Computation, Vol.19 (12), pp 3335-3355 December 2007

158. Scott B B, Constantinople C M, Akrami A, Hanks T H , Brody C D, Tank D W (2017), "Fronto-parietal Cortical Circuits Encode Accumulated Evidence with a Diversity of Timescales", Neuron, Vol.9 (2), pp 385-398, July 2017
159. Seilheimer R L, Rosenberg A, Angelaki D E, (2014), "Models and processes of multisensory cue combination", Current Opinion in Neurobiology, Vol.25, pp 38-46, April 2014
160. Sereno M, Huang R S (2014), "Multisensory maps in parietal cortex", Current Opinion in Neurobiology Vol.24, pp 39–46, 2014
161. Sezgin M C, Gunse B, Kurt G K, (2012), "Perceptual audio features for emotion detection", Journal on Audio, Speech, and Music Processing, Vol 16, December 2012
162. Simoncelli E. P., (2009), "Optimal estimation in sensory systems", The Cognitive Neurosciences, IV, Chapter 36, pp. 525-535, Editor: M. Gazzaniga, MIT Press, 2009
163. Shams L, Ma W J, Beierholm U (2005), "Sound-induced flash illusion as an optimal percept", Neuroreport,Vol.16, Issue 17, pp 1923-1927, November 2005
164. Shams L (2012), "Early Integration and Bayesian Causal Inference in Multisensory Perception", the Neural Bases of Multisensory Processes. Editors: Murray M and Wallace M T, CRC Press, NY, 2012
165. Shomstein S (2012), "Cognitive functions of the posterior parietal cortex: top-down and bottom-up attentional control", Frontiers in Integrative Neuroscience, Vol.6 (38), July 2012
166. Smith A T, Wall M (2008), "Sensitivity of human visual cortical areas to the stereoscopic depth of a moving stimulus", Journal of Vision, Vol.8 (1), August 2008
167. Soltani A, Wang X J (2010), "Synaptic computation underlying probabilistic inference", Nature Neuroscience, Vol.13 (1), pp 112–119, January 2010
168. Steffens B, (2006), "Ibn al-Haytham: First Scientist", Chapter 5. Morgan Reynolds Publishing, Greensboro, North Carolina, USA, 2006
169. Stein B E, Stanford T R, Rowland B A (2014), "Development of multisensory integration from the perspective of the individual neuron" Nature Reviews Neuroscience, Vol 15, pp 520–535, July 2014
170. Stocker A A, Simoncelli E P, (2006), "Noise Characteristics and Prior Expectations in Human Visual Speed Perception", Nature Neuroscience, Vol.9, pp 578-585, 2006
171. Stuphorn V (2016), "Hitting an uncertain target", eLife, 2016 (5), e18721, July 2016
172. Soman S, Jayadeva, Suri M (2016), "Recent trends in neuromorphic engineering" Big Data Analytics, Vol.1, pp 1-15, December 2016
173. Syed I B, (1981), "Islamic medicine: 1000 years ahead of its times", Journal of the Islamic Medical Association of North America, Vol 13, No 1, 1981
174. Szczepanski S M, Konen C S, and Kastner S (2010), "Mechanisms of Spatial Attention Control in Frontal and Parietal Cortex ", The Journal of Neuroscience, Vol.30 (1), pp 148–160, January 2010
175. Theodore M, (2004), "Masters of the Mind: Exploring the Story of Mental Illness from Ancient Times to the New Millennium", John Wiley & Sons, USA, Aug 2004
176. Trappenberg T P (2010), "Fundamentals of Computational Neuroscience", 2nd edition, Oxford university Press, UK, January 2010

177. Stanford T R, Quessy S, Stein B E (2005), "Evaluating the Operations Underlying Multisensory Integration in the Cat Superior Colliculus" Journal of Neuroscience, vol.13 (28), pp 6499-6508, July 2005
178. Triesch J, Von der Malsburg C (2001), "Democratic integration: self-organized integration of adaptive cues", Neural Computing, Vol.13, Issue. 9, pp 2049-2074, September 2001
179. Ursino, M, Magosso E, Cuppini C, (2011), "Sensory Fusion, Perception-Action Cycle, Models, Architectures, and Hardware". 1st edition, Springer, 2011
180. Ursino M, Cuppini C, Magosso E (2014), "Neurocomputational approaches to modelling multisensory integration in the brain: a review", Neural Networks Vol.60, pp 141-165, December 2014
181. Uwe J I (2008), "The role of areas MT and MST in coding of visual motion underlying the execution of smooth pursuit", Vision Research, Vol.48, pp 2062–2069, 2008
182. Ventroux N, Schmit R, Pasquet F, Viel P E, Guyetant S (2009), "Stereovision-based 3D obstacle detection for automotive safety driving assistance", 12th International IEEE Conference on Intelligent Transportation Systems, pp 1-6, October 2009
183. Vingerhoets G (2014), "Contribution of the posterior parietal cortex in reaching, grasping, and using objects and tools", Frontiers in Psychology, Vol.5 (151), March 2014
184. Von Helmholtz, H, (1962). "Treatise on physiological optics", South all J.P.C (Ed), Dover Publications, New York, 1962
185. Wallace M T, Stein B E (1997), "Development of multisensory neurons and multisensory integration in cat superior colliculus", Journal of Neuroscience, Vol. 17 (7), pp 2429-2444, April 1997
186. Wallace M T, Roberson G E, Hairston W D, Stein B E, Vaughan J W, Schirillo J A (2004), "Unifying multisensory signals across time and space", Experimental Brain Research, Vol.158 (2), pp 252-258, April 2004
187. Ward R, Arend I (2012), "Feature binding across different visual dimensions", Attention, Perception, & Psychophysics, Vol.74 (7), pp 1406-1415, October 2012
188. Waniek N, Bremer S, Conradt J (2014), "Real-Time Anomaly Detection with a Growing Neural Gas" 26th International Conference on Artificial Neural Networks (ICANN), Hamburg, Germany, Sept 2014, pp 97-104, September 2014
189. Watkins S, Shams L, Tanaka S, Haynes J D, Rees G (2006), "Sound alters activity in human V1 in association with illusory visual perception", Neuroimaging, Vol.31, Issue 3, pp 1247–1256, July 2006
190. Wei K, Körding K P, (2011), "Causal Inference in Sensorimotor Learning and Control" in Sensory Cue Integration, edited by: Trommershaeuser J, Körding K P, Landy M S, ISBN-13: 9780195387247, Oxford University Press, New York, 2011
191. Wei X X, Stocker A A (2012), "Bayesian Inference with Efficient Neural Population Codes" International Conference on Artificial Neural Networks 2012, Lausanne, Switzerland, pp 523-530, September 2012

192. Weikersdorfer D, Adrian D B, Cremers D, Conradt J (2014), "Event-based 3D SLAM with a depth-augmented dynamic vision sensor" IEEE International Conference on Robotic and Automation, Hong Kong, pp 359 – 364, June 2014
193. Weisswange T H, Rothkopf C A, Rodemann T, Triesch J (2011), "Bayesian Cue Integration as a Developmental Outcome of Reward Mediated Learning", PLOS One, July, 2011
194. Wen B, Boahen K (2009), "A silicon cochlea with active coupling", IEEE Transaction on Biomedical Circuits and Systems, Vol.3, Issue 3, pp.444-455, 2009
195. Wolpert D M, Ghahramani Z, (2000) "Computational principles of movement neuroscience", Nature Neuroscience, Vol 3, pp 1212-1217, November 2000
196. Womelsdorf T, Erxleben K A, and Treue S (2008), "Receptive Field Shift and Shrinkage in Macaque Middle Temporal Area through Attentional Gain Modulation", Journal of Neuroscience, Vol.28 (36), pp 8934-8944, Sept 2008
197. Wozny D R, Beierholm U R, Shams L (2008), "Human trimodal perception follows optimal statistical inference", Journal of Vision, Vol.8, Issue 24, pp 1-11, March 2008
198. Wozny D R, Beierholm U R, Shams L (2010), "Probability Matching as a Computational Strategy Used in Perception", PLOS Computational Biology, August 2010
199. Wright W G, Glasauer S (2006), "Subjective somatosensory vertical during dynamic tilt is dependent on task, inertial condition, and multisensory concordance", Experimental Brain Research, Vol.172 (3), pp 310–321, July 2006,
200. Yang T, Shadlen M N, (2007), "Probabilistic reasoning by neurons", nature, Vol. 447, pp 1075-1082, June 2007
201. Yau J M, DeAngelis G C, and Angelaki D E (2015), "Dissecting neural circuits for multisensory integration and crossmodal processing", Philosophical Transactions of the Royal Society B – Biological Science, 370, September 2015
202. Xu J, Yu L, Rowland B A, Stanford T R, Stein B E (2012), "Incorporating Cross-Modal Statistics in the development and maintenance of multisensory integration ", Journal of Neuroscience, Vol.32, pp 2287-2298, 2012
203. Zimmermann H G, Neuneier R (2000), "Modeling Dynamical Systems By Recurrent Neural Networks", Data Mining II - WIT Transactions on Information and Communication Technologies, Vol 25, 2000
204. Zhang W, Chen A, Rasch M J, and Wu S (2016), "Decentralized Multisensory Information Integration in Neural Systems", Journal of Neuroscience, Vol.13 (2), pp 532-547, January 2016
205. Zhou Y D, Fuster J M (2005), "Somatosensory cell response to an auditory cue in a haptic memory task, Behavioral Brain Research, Vol.153, Issue 2, pp 573–578, August 2004
206. Zhu Y, Quian N (1996), "Binocular receptive fields, disparity tuning and characteristic disparity", Neural Computing, Vol.8, pp 1647-1677, 1996

Eidesstattliche Versicherung / Affidavit

Eidesstattliche Versicherung/Affidavit Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation "*Sensor Fusion in Distributed Cortical Circuits*" selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation "*Sensor Fusion in Distributed Cortical Circuits*" is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

München, October 1, 2019

Mohsen Firouzi

Proof of author contributions

- In Chapter 3, the development of the model, data analysis, programing, designing the experimental setup, and writing the manuscript are done by Mohsen Firouzi. Supervision of the project, hardware and software development of the embedded silicon retina are done by Jörg Conradt.
- In Section 4.1, the main idea, the developed model and data analysis, writing the manuscript, and supervising the project are done by Mohsen Firouzi. The main parts of programming are done by Ahmet Burakhan Koyuncu. The whole manuscript is reviewed by Stefan Glasauer.
- In Section 4.2, the model is developed and analyzed by Mohsen Firouzi, the robotic setup is designed by NST-TUM, and the manuscript is written by Mohsen Firouzi. The project is supervised by Jörg Conradt and Stefan Glasauer.
- The neural model of Chapter 5 is developed and analyzed by Mohsen Firouzi. The project is supervised by Jörg Conradt and Stefan Glasauer, and they helped in interpretation of the results and revision of the manuscript.

Munich, October 1, 2019