Intelligent Holographic Databases

Thesis by

George Barbastathis

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1998

(Submitted October 14, 1997)

To my Family

Acknowledgements

Professor Demetri Psaltis, my graduate advisor, has been an invaluable teacher and friend throughout the defining period of graduate school. The most important thing Demetri taught me was to be uncompromising and swift when proposing, conducting, and presenting scientific research. Demetri has always been present to support me when I was in doubt, criticize me when I erred, and encourage me when I needed it, in both professional and personal matters. I cannot thank him enough for all he did for me during more than four years, and hope to prove worthy of his expectations.

Caltech has been for me a fantastic environment to work and mature scientifically, and that, to a great part, has been due to the interaction and collaboration with distinguished scientists and colleagues. I am grateful to Professors Christof Koch, Yaser S. Abu-Mostafa, Pietro Perona, Shuki Bruck, Richard Andersen, Shin Shimojo, Amnon Yariv, Axel Scherer, Joel Franklin, and Gerry Whitham for letting me share their knowledge and thinking. In particular I thank Christof for his stimulating participation in the "awareness," or "desert survival" project, presented in chapter 7 of this thesis.

Several past and present members of Caltech's Optical Information Processing Group, headed by Dr. Psaltis, contributed directly as collaborators in projects that I worked on. I will never forget the late Ajay Chugh for his gentle, warm personality and mature scientific attitude while he was doing his senior thesis on Shift Multiplexing at Caltech, and for his later help on various collaborations while he was working for Holoplex. I am indebted to Dr. Yong Qiao for teaching me the fundamentals of optical experiments and getting me started with the Electrical Fixing project (section 6.1.1); to Dr. Allen Pu, Michael Levene, and Dr. Kevin Curtis for their contributions to the Shift Multiplexing project (chapters 3, 4); to Dr. Jean-Jacques P. Drolet, Ernest Chuang, and Wenhai Liu with whom I collaborated in the Compact Holographic Memories project (section 6.2), and to Dr. Geoffrey Burr, who

participated in the early development of my ideas about noise effects on holographic memories (chapter 2). Last but not least, I am indebted to Christophe Moser, a close collaborator and friend; together we shared some exciting moments while developping a conoscopic surface topographer, which did not make it into this thesis for lack of coherence with the remaining topics covered herein.

The arguments during group meetings and work in the computer room have been extremely useful. I am grateful to Dr. Hsin-Yu (Sid) Li, Dr. Annette Grot, Dr. Jiafu Luo, Dr. David S. Marx, Robert Denkewalter, Xin An, Ali Adibi, Greg Billock, Greg Steckman, and Xu Wang for their help. Special thanks go to Greg and Greg, my proofreaders. Dr. Kelvin Wagner, Dr. Ju-Seog Jiang, and Dr. Chuan Xie were great sources of stimulating discussions and inspiration, and good friends. If there has been a single key factor in the success of whatever experiments I have done, that was our technician Yayun Liu. Lucinda Acosta, our group's administrative assistant, and Linda Dosza, Iliana Salazar, Janette Aguado, Helen Carrier, and Lynn Hein have been patient enough to cope with my incompetence when it comes to formal paperwork. Paula Samazan used to provide a depository of wisdom in her small EE/APh library, which unfortunately does not exist anymore. I am also grateful to Tanya Erdmann of Sponsored Research for some terrific help during last-minute proposal preparations, and to Dr. Dean Schonfeld, the ERC Technology Transfer officer, for a lot of stimulating discussions about applied research and the future of Neuromorphic Engineering. Jean-Jacques, Bob Freeman and David L. Sieving deserve a great "thanks" for being my computer mentors and saving me in several occasions from the deadly consequences of computer glitches or my own computer illiteracy.

Some of my best grad school memories are from mountains and beaches in California and Mexico, at clubs and bowling arcades in Los Angeles and the Bay Area, at "HeadQuarters" parties and Tournament Park asados with my friends Wayez Ahmad (the "Commander"), Jean-Yves Bouguet, Luis Goncalves, Mario and Pili Munich, Enrico and Barbara Di Bernardo, Francesco Bullo, Costas Christoyiannis, Alberto Pesavento, Diego Dugatkin, Gudrun Socher, Dieter Koller, Melissa Saenz, Laurent Chognard (the "Barbarian"), and Arrigo Benedetti. They undertook the enormous

task of getting me out of the shell once in a while; I am grateful for they succeeded the right amount – no more, no less! Polly Preventza, my long-time peer since college, has been a great partner in hunting for sushi restaurants and speaking Greek and Spanish with me when English was getting frustrating.

Outside Caltech, several people have influenced my shaping as a scientist. My deepest gratitude goes to my advisors at the National Technical University of Athens, Professors Christos N. Capsalis and Nikolaos K. Uzunoglu, for guiding me through my first research steps. Among all the wonderful teachers and friends I had at NTUA, I will always be grateful to Professors John G. Fikoris and Yiannis Tsividis for their principles and inspiration. I have been fortunate enough to work with distinguished researchers at many other places: Professor Brian H. Kolner at the University of California, Los Angeles (now with UC, Davis) and his students Corey Bennett, Ryan Scott and Bill Stanton; Dr. John H. Hong, Dr. Tallis Y. Chang, Dr. Jian Ma and Dr. Ratnakar R. (R^2) Neurgaonkar at the Rockwell International Science Center; Professor Peter J. McDonnell at the University of Southern California; and Dr. Gan Zhou and Dr. Fai Mok at Holoplex. I am indebted to all for their continued help and friendship. It is difficult for me to express in simple words my respect and admiration for Professor Nicolaos G. Alexopoulos, who undoubtedly had the most influence in the choices I made in life and continues to provide me with advice and inspiration in his unparalleled calm and friendly manner.

Even though in different scientific fields, my parents, grandparents and family deserve the credit for orienting me towards excellence, and teaching me the relentless pursuit of the truth. They patiently helped me during the first difficult steps of learning; they were present to support me when tough decisions were demanded; they saught to provide me the difficult qualities of culture and principle. For their unassuming love, I cannot thank them enough.

Abstract

Memory is a key component of intelligence. In the human brain, physical structure and functionality jointly provide diverse memory modalities at multiple time scales. How could we engineer artificial memories with similar faculties? In this thesis, we attack both hardware and algorithmic aspects of this problem.

A good part is devoted to holographic memory architectures, because they meet high capacity and parallelism requirements. We develop and fully characterize shift multiplexing, a novel storage method that simplifies disk head design for holographic disks. We develop and optimize the design of compact refreshable holographic random access memories, showing several ways that 1 Tbit can be stored holographically in volume less than 1 m^3, with surface density more than 20 times higher than conventional silicon DRAM integrated circuits. To address the issue of photorefractive volatility, we further develop the two-lambda (dual wavelength) method for shift multiplexing, and combine electrical fixing with angle multiplexing to demonstrate 1,000 multiplexed fixed holograms. Finally, we propose a noise model and an information theoretic metric to optimize the imaging system of a holographic memory, in terms of storage density and error rate.

Motivated by the problem of interfacing sensors and memories to a complex system with limited computational resources, we construct a computer game of *Desert Survival*, built as a high-dimensional non-stationary virtual environment in a competitive setting. The efficacy of episodic learning, implemented as a reinforced Nearest Neighbor scheme, and the probability of winning against a control opponent improve significantly by concentrating the algorithmic effort to the virtual desert neighborhood that emerges as most significant at any time. The generalized computational model combines the autonomous neural network and von Neumann paradigms through a compact, dynamic central representation, which contains the most salient features of the sensory inputs, fused with relevant recollections, reminiscent of the hypothesized

cognitive function of awareness. The Declarative Memory is searched both by content and address, suggesting a holographic implementation. The proposed computer architecture may lead to a novel paradigm that solves "hard" cognitive problems at low cost.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Memory and intelligence

Memory... You gave them memories!

R. Deckard in *Blade Runner*

What is the significance of memory for human intelligence? Our recollections influence and enrich our behavior, because they bring experience to our assistance. It is thanks to the amazingly successful organization of human memory that highly complex cognitive tasks, e.g., face recognition, are possible. On the other hand, this same memory is deficient for most humans in other types of tasks, such as memorizing long lists of objects. Understanding human memory organization will probably help treatment and prevention of memory defects. Another exciting possibility is mimicking human memories in the construction of artificial memories for intelligent machines. This may result in systems capable of exhibiting, to some extent, the richness and adaptability of human behavior.

One tends to think of memory as a unified function; however, amnesic patients are typically deprived of only a limited class of tasks, while maintaining dexterity in others [1, 2, 3, 4]. This indicates that human memory is organized in specialized systems and in a highly distributed manner. To appreciate the varsity of human memory storage, compare learning how to drive a car, memorizing a phone-number, and understanding how to solve differential equations. Conscious experiences sometimes enter memory with surprising clarity and stay there for a long time, while others pass without leaving a trace. There is experimental evidence that the converse is also true: memories of rather complex relations between a sequence of experiences can be formed and used without the participation of consciousness [5, 6].

Physiological research on memory started in the early 1880's [1, 2, 7, 8]; however,

it was not until the late 1970's-early 1980's that definite evidence on the multiple memory modalities was obtained from psychophysical experiments in normal subjects and patients with lesions [9]. The human memory systems accepted by researchers today may be classified in six major categories [3, 4, 9, 10, 11]:

1. Iconic/echoic, which are pre-categorical, fast decaying visual/auditory memories and can be thought of as scratch-space for perception [12];
2. Short-term, containing very recent cognitive information which needs to be easily accessible [13];
3. Perceptual, involved in priming of identification of objects;
4. Procedural, used in skill learning, and simple conditioning;
5. Semantic, representing general factual knowledge; and
6. Episodic, storing factual recollections (events) from the personal past.

Episodic and Semantic memories are sometimes coallesced into the single term Declarative Memory because they have to do with cognitive information, while the rest are termed Non-Declarative. In the literature there is no specific information about which subsystems contain consciously formed memories, even though it seems safe to say that Episodic memories are formed consciously, whereas Semantic memory may be unconscious (it is also often classified as "implicit" memory along with Procedural memory).

Examining the "hardware" of human memory, we find that among the brain areas involved in memory formation, the hippocampus is activated during the formation of flexible memories, i.e., memories that must be extended to situations substantially different than the original memorized event. In these cases, the useful information results from the *relations* between different memory objects. Evidence towards that effect has been obtained for humans and animals. For example, it has been shown [9, 14, 15] that rats with hippocampal damage are able to swim to a platform that they have learnt previously only if they start at the same location as in the learning trial. They fail if the starting position is different.

The hippocampus is also activated during the presentation of novel stimuli. For

instance, in a task of detecting unknown sentences compared to a sequence learnt 24 hours earlier, limbic structures such as hippocampus, parahippocampal gyrus, medial dorsal thalamus, and medial frontal cortex were found to modify their activity in response to novelty [16]. In a reaction time task, where performance was improving as result of implicit learning of a finite grammar by the subjects, different areas were activated: the right ventral striatum responds to novelty, whereas the right prefrontal area maintains contextual information used for the task [6]. The distinction is very important, because it indicates that memory formation in some parts of the brain is conscious whereas in other parts it is not.

There is at least one more occurrence of the prefrontal cortical areas patently participating in unconscious learning. In a gambling game, where participants did not know the risks a priori, it was observed that normal subjects were able to improve their performance and generate anticipatory skin conductance responses (SCR's) before they became consciously aware of the risk, whereas patients with prefrontal lesions were unable to either learn or generate SCR's [5]. Since the prefrontal cortex is also strongly involved in macaque monkeys in motor planning together with the posterior parietal cortex (PPC) [17], it might be the case that the prefrontal cortex becomes activated in preparation for other brain areas (the PPC in the case of motor planning, the ventral striatum in the case of novelty detection in implicitly learnt facts, and possibly the hippocampus in the case of explicit memory – see below), thus acting as a sort of CPU that assigns tasks to other brain areas.

The significance of attention in the distinction between the two types of learning is clear. In the case of conscious detection, the subject's attention is devoted to the detection task, whereas when learning is implicitly utilized, the subject is performing a seemingly unrelated reaction task. The presence or absence of attention also determines the time course of perceptual learning. During the course of learning a texture discrimination skill, when attention is of course focused to the task, subject performance saturates; however, after several hours subject performance is found to improve significantly, indicating latent "consolidating" changes in the primary visual cortex [18, 19].

This type of implicit learning is perhaps associated with the transfer of knowledge between different memory systems. For example, during the learning phase of a new motor task, such as playing tennis, most of the conscious effort is spent on controlling the muscles so that they follow the trainer's instructions, which correspond to specific events. Therefore, Declarative Memory must be in use during that learning phase. Later, after training has progressed enough, muscle control is automatic in response to more complicated considerations, such as trying to predict the opponent's next position. For an experienced player, persistence in learning apparently causes the motor aspects of tennis playing to transfuse into Procedural Memory.

The understanding of human memory is still far from assigning specific neuronal mechanisms to effects such as transfusion between sub-systems. Neither do artificial memories have any similar capability, which may be the reason why computers are currently inefficient in many cognitive tasks (e.g., face recognition or chess[1]).

Some form of memory organization is present in most computers. For example, a personal computer (PC) has the following hierarchy:

1. Permanent memory (hard-drive, CD-ROM), contains data that are infrequently or never changed, and trades off a small overhead in retrieval time (typically 10 nsec/bit) for large capacity (typically several GBytes).
2. Random-access memory (RAM), the computer's working memory where data are continuously stored and modified by the active programs; it is critical that the access time to the RAM is small (typically 10 nsec/byte), but despite that silicon RAM's as large as 1-2 GBytes are nowadays available.
3. Cache memory, containing copies of the most recently accessed RAM data, making them available to the processor for fast re-access; access time is smaller than the RAM (typically 1 nsec/word) and the capacity is small (256-512 kbytes).

[1]In May 1997, Garry K. Kasparov, World Chess Champion, lost a match $2^1/_2 - 3^1/_2$ to *Deep Blue*, a massively parallel super-computer constructed by IBM. Until that match, chess-playing had been considered an exemplary "hard" cognitive task, because human superiority over computers had been customary at the professional level. Thanks to its superior hardware, at each move *Deep Blue* exhaustively considered possible continuations to great depth sufficiently fast to compete successfully within the allotted time limit. Therefore, the match result shows that chess is actually amenable to "exhaustive calculation." IBM has since discontinued the *Deep Blue* project.

The memory hierarchy of the PC is diverse in hardware and interconnection between modules (e.g., cache co-resides with the processor, whereas RAM and hard-drive are separate modules). However, the hardware does not have the built-in capability of selectively transfusing stored salient memories between different sub-systems as happens in the higher primates.

While the demand for ever powerful computers continues to increase, a novel desire for interactive machines is surfacing in various domains. For instance, in industry, it is desirable to use intelligent robots for work in hazardous environments; in multimedia or virtual reality applications, the computer should be able to "interpret" the user's behavior using traditional input devices (mouses, keyboards) as little as possible. In order to demonstrate such levels of computer intelligence, rather than continuing the current trend of research for incremental improvements in processors and algorithms, it may be more rewarding to seek novel maximum-efficiency computational architectures where the data organization (the "memory") is tailored to the demands of the required applications. This approach is evidently successul in biological systems, culminating with the human brain.

The focus of this thesis is on constructing memories for intelligent machines. We focus on novel architectures for holographic memories as hardware solution, because they have two properties that are apparently present in human memories as well: (1) they are parallel[2], (2) they allow data to be retrieved associatively. Moreover, direct recall (by address) is also possible in holographic memories. From the algorithmic viewpoint, we are interested in how multiple sensory modalities may generate representations appropriate for efficient storage. We explore this theme in Chapter 7, having in mind the holographic implementation and its implications.

[2]For holographic memories, the correct term is "page-oriented." See the next section for a more detailed introduction.

1.2 Holographic databases

Holographic memories [20, 21, 22] are particularly suitable for applications that demand high data capacity, high transfer rate and the presence of both direct and associative recall. Holography was invented by Gabor in 1948 [23, 24, 25], and volume holography was proposed as a method for data storage as early as 1963 [20, 26]; however, early efforts [27, 28] did not indicate its viability for large-scale production. This opinion was reversed in the early 90's when Fai Mok demonstrated that it was possible to store an unprecedented amount of information holographically [29]. More recently, the Holographic Random Access Memory [30, 31] (HRAM) and the three-dimensional (3-D) Holographic Disk [32, 33] architectures were proposed and demonstrated for high capacity digital storage [22, 27, 34]. These achievements were enabled largely because of several reasons: *(i)* significant improvements in optoelectronic devices (spatial light modulators, CCD cameras) that were used to interface the memory with a computer for writing and recording information, *(ii)* progress in the understanding of the dynamics of well-known materials such as photorefractives [35], *(iii)* the appearance of new materials such as the photopolymers. In turn, the encouraging results sparked increased activity in the field of holographic storage. In this thesis we will present several novel architectures, aimed at improved density, access time, and portability for holographic databases.

Two typical configurations used in holographic memory systems are the transmission and 90° (90-degree) geometries, shown in Figures 1.1 and 1.2 respectively. The reference is a plane wave. In the signal arm, a Spatial Light Modulator (SLM) imprints the information on the wavefront (usually as amplitude modulation), which is then recorded as a hologram by interfering with the reference. Multiplexing is achieved by changing the angle of incidence of the reference (angle-multiplexing [26, 36]) using the mirror and the telescopic 4-F system. The reconstruction is imaged on the CCD camera for detection. Except for angle multiplexing, a number of different multiplexing techniques have been proposed for holographic storage, such as wavelength [37, 38], phase-code [39], fractal [40], and peristrophic [41]. In Chapters 3 and

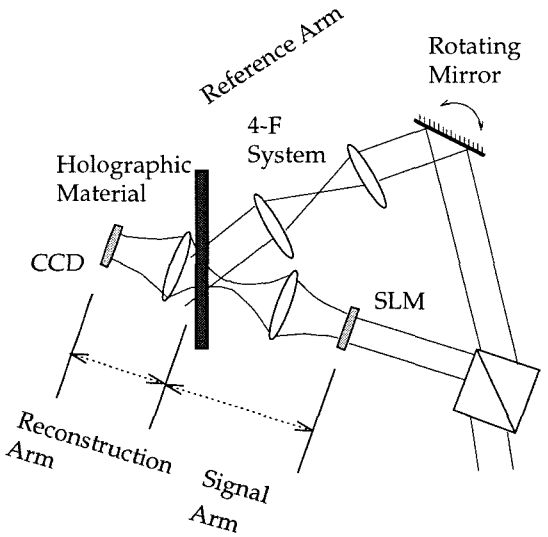


Figure 1.1: Angle-multiplexed holographic memory in the transmission geometry.

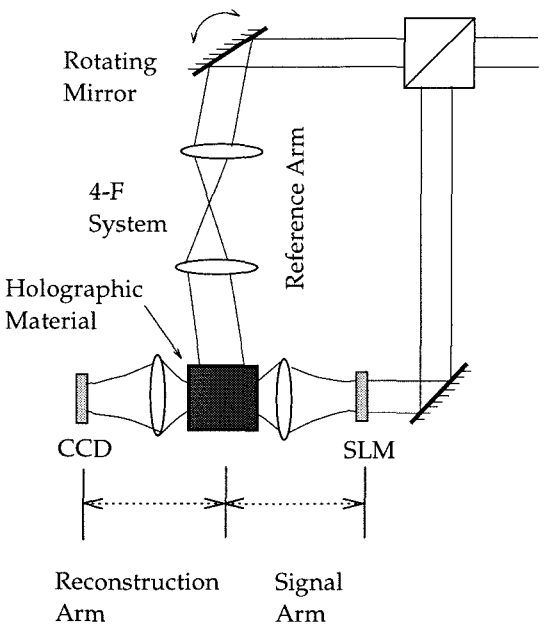


Figure 1.2: Angle-multiplexed holographic memory in the 90° geometry.

4 we will describe in detail two novel methods: array and shift multiplexing [42, 43].

Transmission geometry, as in Figure 1.1, is used in the holographic 3D disk [32, 33, 44] architecture, illustrated in Fig. 1.3. High capacity is achieved if two multiplexing methods combined (e.g., angle and peristrophic [33]) are used to superimpose holograms on the same location (in Fig. 1.3 the angle multiplexing mechanism alone is shown). After the total number of holograms allowed by the geometry and the medium dynamic range is used up, rotational motion of the disk is utilized to access different locations on the disk surface where the process is repeated (spatial multiplexing). Typically, the size of each location containing multiple holograms is a few square millimeters. This poses a challenge in the implementation, since disk motion cannot be continuous during recording or readout; it rather occurs in the form of "jumps" from one location to the other. Alternatively, a continuously spinning disk can be used, but the light source needs then to be pulsed. In most multiplexing techniques, some mechanism is needed inside the disk head in order to implement selective read-out. Angle and peristrophic multiplexing require a beam steering mechanism; wavelength multiplexing requires an accurate, high resolution tunable laser source;

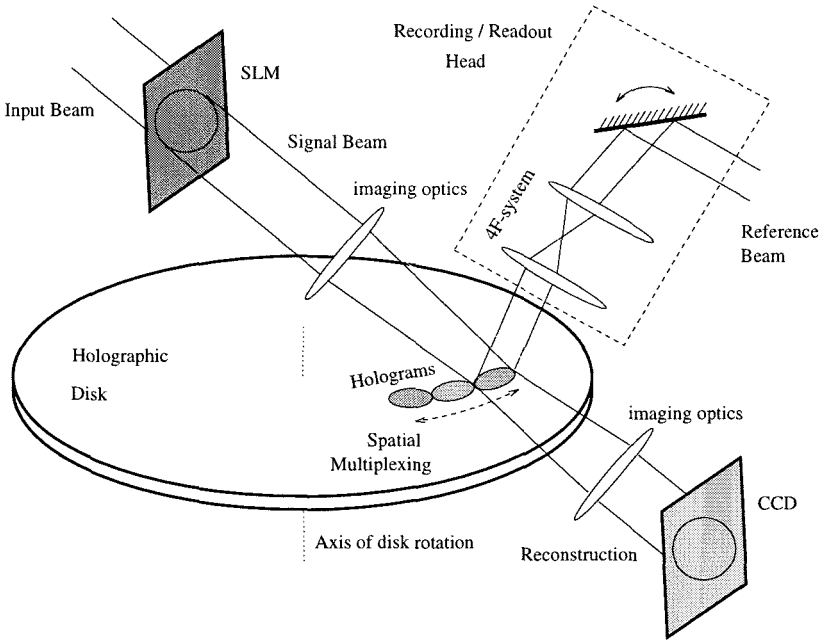


Figure 1.3: Design of a 3–D holographic disk.

phase-code multiplexing requires a second SLM to implement the orthogonal phase code used for reference. In the cases of array and shift multiplexing, access is easier, because different holograms become Bragg-matched by utilizing disk rotation alone, as we will describe in detail later.

If the reference and signal beams illuminate orthogonal faces of the holographic medium, the 90°-geometry variant of Fig. 1.2 is obtained. In this architecture, the angular selectivity is very small, allowing very dense packing of the holograms, although this occurs at the expense of dynamic range. The 90°-geometry has been widely used in high capacity HRAM demonstrations [29, 34, 45, 46]. Again, it is necessary to combine spatial multiplexing with one or two other techniques (usually angle+fractal [45, 46]) in order to maximize the capacity.

An important design decision concerns the holographic material. For read-only (or write-once-read-many, WORM) applications, photopolymers like DuPont's HRF-100 are promising [33, 47]. On the other hand, rewrittable holographic memories [48] are usually implemented with photorefractive materials [49, 50, 51, 52]. Rewrittable

holographic memories are volatile during readout, and even in the dark (i.e., the stored information decays even when it is not being read-out, as is the case for CMOS memories). We will discuss several solutions to this problem in Chapter 6. Other possible materials include photorefractive polymers [53, 54, 55], but these are still in the early development phase in terms of the understanding of the physics of recording, and of practical applications.

Holographic memories are peculiar in the sense that the minimum retrievable unit is a page (from several Kbits to a few Mbits) rather than a single bit. For direct access, parallelism is usually an advantage, unless the data within a page are segmented in small records (of a few bytes each) because then the additional processing needed to extract the individual records from the page may slow down the memory. Additional issues are the error rate, the coding scheme and the memory interface. Several of the trade-offs between memory density, error rate and access method will be discussed in Chapters 2, 5, and 6.

1.3 Outline of the thesis

We start by developing the mathematical framework for noise in readout from page-oriented optical memories in Chapter 2. In particular, we are concerned with the trade-off of pixel size and storage density against bit error rate, and use the measure of information capacity to derive the upper bound on the useful information that can be stored in a page-oriented memory.

A novel holographic storage method, using non-planar reference waves, is fully developed and analyzed in Chapters 3 and 4. We consider two types of reference beams: first a fan of plane waves ("array" multiplexing, section 3.2), and then a single spherical wave ("shift" multiplexing, section 3.3). Shift multiplexing is of particular interest because it is easy to implement and has already been used in experimental high capacity demonstrations. We devote the entire Chapter 4 on the performance of shift-multiplexed memories: crosstalk (section 4.1), dynamic range (section 4.2), image distortion (section 4.3), and surface storage density (section 4.4) for shift-

multiplexed holographic 3D disks, and finally nonvolatile readout of photorefractive shift-multiplexed memories (section 4.5).

In Chapter 5 we are concerned with the problem of choosing an imaging system for a high capacity memory (using surface storage density as metric). We consider the 4–F (in the Fourier and image plane storage geometries) and van der Lugt imaging systems and analyze the constraints they impose on imaging in section 5.1. We present in some detail the main derivations for the simplest and best-known case, angle multiplexing in section 5.2. We then provide the main results for two more multiplexing methods: wavelength (section 5.3) and shift (section 5.4), in the latter case using the theory developed in Chapters 3 and 4.

Chapter 6 addresses several other issues in holographic memory design, in particular volatility (section 6.1, with emphasis on electrical fixing, section 6.1.1), compactness and refreshing, with emphasis on the volume and error rate of a refreshable Tbit HRAM that physically occupies less than 1 m^3 (section 6.2), and associative access, with emphasis on shift-multiplexed and compact architectures (section 6.3).

In Chapter 7 we present a computer architecture appropriate for adaptive intelligent systems. The two most interesting aspects in the design, from the point of view of this thesis are (i) the central concept of "awareness," a computational bottleneck that allows the main processor to operate in real time by reducing its input space dimensionality in a dynamic, adaptive fashion; (ii) the declarative memory, which should provide both direct and associative access; hence, a holographic memory comes to mind as a good candidate for its implementation. The motivation for the awareness approach to computation came from a computer game of *Desert Survival*, described in section 7.1. The general-purpose awareness-based architecture and its similarities to physiological and psychological models are presented in section 7.2. The memory organization and dynamics are studied in section 7.3; ideas on how holographic memories may be be fused into the field more effectively are given in section 7.3.1, and the "reinforced Nearest Neighbor" (rNN) learning scheme is introduced and analyzed in sections 7.3.2-7.3.6. The concluding section 7.4 discusses applications of the computational paradigm of Chapter 7, and future directions in related research.

Chapter 2 Noise in page-oriented optical memories

2.1 Introduction

Intensity detection is used almost exclusively in practical continuous wave optical systems, because it is by far the cheapest and easiest method. However, the phase information of the optical signals is lost during the intensity-forming operation. For example, if the signal is real and the noise contains an imaginary component, then the imaginary noise will affect the performance of an intensity-detecting system, but not that of an interferometric (phase-detecting) system. On the other hand, interferometric systems suffer from severe sensitivity to all kinds of environmental disturbances; the effect of these on signal quality is so bad that the benefits from extracting phase information are lost.

In material presented later in this thesis, the noise performance of an optical memory is important. Therefore, here we concentrate on the analysis of intensity detecting systems. In particular, we consider digital systems affected by complex Gaussian noise, and analyze the error performance, using the common metric of bit error-rate (BER), a fancy way of saying "probability of making an error in the detection of a single bit." We also discuss the concept of signal-to-noise ratio (SNR) in the context of a digital system, describe the properties one would require of such a quantity, and define a metric that satisfies them in the context of intensity detecting systems.

An important feature of optical detectors is *integration*, meaning the *incoherent* addition of elementary electric currents induced by the incident intensity. If these elementary currents are statistically independent to some extent, we expect that the integration improves error performance. As a quantitative measure of the degree

of independence, we extend the notion of "effective degrees of freedom" [56] to the space domain and to stochastic processes with non-zero mean, and indicate how to measure it experimentally. Statistical independence increases with the surface of the detector, but at the expense of surface density. This trade-off is treated with the aid of an information density measure in section 2.3 and is the concluding result of this chapter.

A legitimate question is: why bother deriving the statistics of the intensity before and after integration? Since the noise contains a large number of components, one may invoke the Central Limit Theorem (CLT), and use Gaussian statistics for the intensity. This is not quite accurate. The CLT of course applies, but is accurate only in the central mass of the distribution [57]; in many cases (intensity signals derived from Gaussian-type coherent signals among them) the tails of the true distribution deviate largely from the CLT approximation. The BER depends on the probability mass at the distribution tails, where the CLT approximation predicts higher BER than our more accurate approach.

Notation for Chapter 2

We will use boldface notation for random variables or stochastic processes. Thus $p_{\mathbf{V}}(v)$, $\mathrm{EV}\{\mathbf{V}\}$, $\mathrm{Var}\{\mathbf{V}\}$ are the probability density function (pdf), expectation value, and variance, respectively, of the random variable $\mathbf{V}$. In a good part of this chapter we will be dealing with estimating the moments of the random variable or stochastic process $\mathbf{V}$ from statistical data; we will denote by $\mathbf{m_V}$ for the statistical estimate of the mean $\mathrm{EV}\{\mathbf{V}\}$, and by $\boldsymbol{\sigma}_{\mathbf{V}}$ the statistical estimate of the standard deviation of $\mathbf{V}$, i.e., $\boldsymbol{\sigma}_{\mathbf{V}}^2 = \mathrm{EV}\{(\mathbf{V} - \mathrm{EV}\{\mathbf{V}\})^2\}$.

All stochastic processes are assumed wide-sense stationary, unless stated otherwise. It will turn out to be notationally convenient to use vector notation for the entire ensemble of a stochastic signal. Thus

$$\mathcal{V} = (\mathbf{V}(x_1) \ldots \mathbf{V}(x_K))^T$$

is a column vector containing the values of the stochastic process $\mathbf{V}$ at points $x_1, \ldots x_K$.

Traditionally, it is more natural to think of a stochastic process as being defined in a continuum. We will avoid considerable complications by sticking to discrete processes, even in the case $K \to \infty$.

The optical signal $\mathbf{E}$ incident on the detector is a stochastic process in the spatial variable x, defined as

$$\mathbf{E}(x) = \mathbf{S} + \mathbf{N}_{\mathrm{r}}(x) + i\mathbf{N}_{\mathrm{i}}(x). \tag{2.1}$$

$\mathbf{S}$ is the signal process, taking values S ("ON" pixel) and 0 ("OFF" pixel) with equal probabilities. $\mathbf{N}_{\mathrm{r}}$, $\mathbf{N}_{\mathrm{i}}$ are the in-phase and quadrature components of the noise, respectively. They are assumed statistically independent identically distributed (iid) Gaussian random processes with mean 0 and variance σ^2; therefore, they both follow the probability density function

$$p_{\mathbf{N}}(n) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{n^2}{2\sigma^2}\right\}. \tag{2.2}$$

The normalized joint moments of the noise processes are assumed symmetric and are denoted as

$$\gamma_{jk}(\xi) = \frac{\mathrm{EV}\left\{\mathbf{N}_\alpha^j(x+\xi)\mathbf{N}_\alpha^k(x)\right\}}{\sigma^{j+k}}, \tag{2.3}$$

where α may be "r" or "i."

The detector senses the intensity process $\mathbf{I}$, which is defined as

$$\mathbf{I}(x) = |\mathbf{E}(x)|^2 = (\mathbf{S}(x) + \mathbf{N}_{\mathrm{r}}(x))^2 + \mathbf{N}_{\mathrm{i}}^2(x). \tag{2.4}$$

We will use the symbol $\gamma_{jk}^{\mathbf{I}}(\xi)$ for the autocorrelation of the intensity process $\mathbf{I}$.

The decision whether the detected pixel is "ON" or "OFF" can be taken either based on the value of $\mathbf{I}$ on a single location x inside the pixel or after spatial integration over the entire pixel. To treat the latter case, we define the integrated intensity $\mathbf{W}$

as

$$\mathbf{W} = \frac{1}{\mu} \sum_{n=1}^{\mu} \mathbf{I}(x_n). \tag{2.5}$$

The quantity μ is very interesting: from a mathematical viewpoint, if we let $\mu \to \infty$, then we obtain the Riemann integral

$$\mathbf{W} \xrightarrow[\mu \to \infty]{} \int \mathbf{I}(x')\mathrm{d}x'. \tag{2.6}$$

This integral is well defined as a random variable, but does not shed light on the physical significance of μ. For our purposes, it suffices to consider μ as the number of elementary detector currents contributing incoherently to the overall current produced by light entering the detector. Thus μ is a very large number, so that it is legitimate to consider it infinite if this turns out to be mathematically convenient; physically it is reassuring to think that it remains finite.

2.2 Intensity detection of signals in Gaussian noise

2.2.1 Statistical properties of optical intensity

Probability density functions

The statistical properties of a random variable are entirely specified if its pdf is known. The amplitude process $|\mathbf{E}|$ follows the Rayleigh distribution for dark pixels and the Rician distribution for bright pixels [58]; therefore, the respective intensity processes are exponential and affine χ^2 (chi-square) with two degrees of freedom[1]:

$$p_{\mathbf{I}}\left(I|\mathbf{S}=0\right) = \frac{1}{2\sigma^2} \exp\left\{-\frac{I}{2\sigma^2}\right\}, \tag{2.7}$$

$$p_{\mathbf{I}}\left(I|\mathbf{S}=S\right) = \frac{1}{2\sigma^2} \exp\left\{-\frac{I+S^2}{2\sigma^2}\right\} \mathrm{I}_0\left(\frac{S\sqrt{I}}{\sigma^2}\right), \tag{2.8}$$

where $\mathrm{I}_\nu(.)$ is the modified Bessel function of 1st kind and order ν. Equation (2.8) is

[1] The two degrees of freedom correspond to the independent processes $\mathbf{N}_\mathrm{r}$ and $\mathbf{N}_\mathrm{i}$.

also called Erlang distribution.

The integrated intensity $\mathbf{W}$ is the summation of the intensities of 2μ Gaussian random variables (see eq. 2.5). If these variables are independent, then the pdf for $\mathbf{W}$ is the μ-fold convolution of individual pdf's of the form (2.7) for dark pixels, and (2.8) for bright pixels. This results in the following expressions:

$$p_{\mathbf{W}}(W|\mathbf{S}=0) = \left(\frac{\mu}{2\sigma^2}\right)^{\mu} \frac{W^{\mu-1}}{\Gamma(\mu)} \exp\left\{-\frac{\mu W}{2\sigma^2}\right\}, \tag{2.9}$$

$$p_{\mathbf{W}}(W|\mathbf{S}=S) = \frac{1}{2\sigma^2}\left(\frac{W}{\mu S^2}\right)^{\frac{\mu-1}{2}} \exp\left\{-\frac{W+\mu S^2}{2\sigma^2}\right\} \mathrm{I}_{\mu-1}\left(\frac{S\sqrt{\mu W}}{\sigma^2}\right), \tag{2.10}$$

where $\Gamma(.)$ is the Gamma function. We recognize (2.10) as the affine χ^2 distribution with 2μ degrees of freedom. As the Riemann integral limit is approached (i.e., $\mu \to \infty$), the distributions tend to become impulse-like, centered at $2\sigma^2$ and $S^2 + 2\sigma^2$ for dark and bright pixels, respectively. The derivation of (2.10) is non-trivial and will be given in Appendix I.

The independence assumption is very convenient for calculations, but holds for white noise only; in practice, the spatially bandlimited nature of spatial optical noise is not negligible, so the approach presented above needs to be modified. We start with the case $\mathbf{S} = 0$ (dark pixel) and consider the real and imaginary noise vectors

$$\mathcal{N}_{\mathrm{r}} = (\mathbf{N}_{\mathrm{r}}(x_1) \ldots \mathbf{N}_{\mathrm{r}}(x_\mu))^T \quad \text{and}$$
$$\mathcal{N}_{\mathrm{i}} = (\mathbf{N}_{\mathrm{i}}(x_1) \ldots \mathbf{N}_{\mathrm{i}}(x_\mu))^T.$$

We may express the joint pdf of the ensemble of random variables $\{\mathcal{N}_{\mathrm{r}}, \mathcal{N}_{\mathrm{i}}\}$ as follows[2]:

$$p_{\mathcal{N}_{\mathrm{r}}\mathcal{N}_{\mathrm{i}}}(\overline{n}_{\mathrm{r}}, \overline{n}_{\mathrm{i}}) = \frac{1}{(2\pi)^{\mu}\,|\Sigma|} \exp\left\{-\frac{1}{2}\overline{n}_{\mathrm{r}}^T \Sigma^{-1} \overline{n}_{\mathrm{r}} - \frac{1}{2}\overline{n}_{\mathrm{i}}^T \Sigma^{-1} \overline{n}_{\mathrm{i}}\right\}, \tag{2.11}$$

[2]Recall that $\mathcal{N}_{\mathrm{r}}$ and $\mathcal{N}_{\mathrm{i}}$ are independent by assumption.

where Σ is the correlation matrix, with entries

$$\Sigma_{mn} = \sigma^2 \gamma_{11}\left(x_m - x_n\right), \tag{2.12}$$

and $|\Sigma|$ is the determinant of Σ. Since Σ is symmetric and positive definite, it must have exactly μ real non-negative eigenvalues λ_j^2 $(j = 1, \ldots, \mu)$. Therefore a unitary transformation Q exists such that

$$\Lambda = Q^T \Sigma\, Q \tag{2.13}$$

is diagonal (with the eigenvalues as elements). If we now let

$$\tilde{\mathcal{N}}_{\mathrm{r}} = Q\, \mathcal{N}_{\mathrm{r}}, \tag{2.14}$$

$$\tilde{\mathcal{N}}_{\mathrm{i}} = Q\, \mathcal{N}_{\mathrm{i}}, \tag{2.15}$$

then the components of $\tilde{\mathcal{N}}_{\mathrm{r}}$, $\tilde{\mathcal{N}}_{\mathrm{i}}$ are also normally distributed and *they are* independent [57]. After substituting the transformation into (2.11) we obtain

$$p_{\tilde{\mathcal{N}}_{\mathrm{r}}\tilde{\mathcal{N}}_{\mathrm{i}}}(\overline{n}_{\mathrm{r}}, \overline{n}_{\mathrm{i}}) = \frac{1}{(2\pi)^{\mu}\,|\Lambda|} \exp\left\{-\frac{1}{2}\overline{n}_{\mathrm{r}}^T \Lambda^{-1} \overline{n}_{\mathrm{r}} - \frac{1}{2}\overline{n}_{\mathrm{i}}^T \Lambda^{-1} \overline{n}_{\mathrm{i}}\right\}. \tag{2.16}$$

Since the above equation is already decoupled into the individual pdf's of the components of vectors $\tilde{\mathcal{N}}_{\mathrm{r}}$, $\tilde{\mathcal{N}}_{\mathrm{i}}$, we can again use the convolution approach to derive the exact pdf of $\mathbf{W}$. The derivation is carried out in Appendix II and yields the result

$$p_{\mathbf{W}}\left(W|\mathbf{S}=0\right) = \sum_{m=0}^{\mu} \frac{\frac{1}{2\lambda_m^2} \exp\left\{-\frac{W}{2\lambda_m^2}\right\}}{\prod_{m' \neq m}\left(1 - \frac{\lambda_{m'}^2}{\lambda_m^2}\right)}. \tag{2.17}$$

In the Riemann integral limit ($\mu \to \infty$), this approach is formally equivalent to a Karhunen-Loève expansion [56], and yields the same result.

This accurate method is expensive computationally, and not very practical if the correlation function is not known exactly. In many common cases, it is reasonable

to assume that only a few out of the M eigenvalues of Σ are appreciably larger than zero. We will denote the number of "large" eigenvalues by μ_{eff} ($\mu_{\text{eff}} < \mu$). The remaining $(\mu - \mu_{\text{eff}})$ components of $\tilde{\mathcal{N}}_{\text{r}}$, $\tilde{\mathcal{N}}_{\text{i}}$ are distributed effectively like strong spikes; therefore, they do not affect the convolution operation significantly. It follows that the pdf of $\mathbf{W}$ is in effect the result of convolving μ_{eff} only independent components. (This statement is exact if the remaining $(\mu - \mu_{\text{eff}})$ eigenvalues are precisely equal to zero [57], because then the spikes become delta-functions). Further simplification follows if we assume that all nonzero eigenvalues are approximately equal to σ^2. This approximate procedure is justified well for optical systems, as we will show in section 2.2.2. The result is the much simplified pdf

$$p_{\mathbf{W}}(W|\mathbf{S}=0) \approx \left(\frac{\mu_{\text{eff}}}{2\sigma^2}\right)^{\mu_{\text{eff}}} \frac{W^{\mu_{\text{eff}}-1}}{\Gamma(\mu_{\text{eff}})} \exp\left\{-\frac{\mu_{\text{eff}} I}{2\sigma^2}\right\}. \tag{2.18}$$

This is the same as (2.9) but with μ_{eff} instead of μ.

We will refer to μ_{eff} as the "effective degrees of freedom" of the variables appearing in summation (2.5). It turns out that if μ_{eff} is correctly estimated, then (2.9) approximates the tails of the exact distribution reasonably well [56]. The reason is that for large W the importance of the omitted terms in (2.17) diminishes, as we will argue in Appendix III.

A similar procedure for the bright pixels ($\mathbf{S} = S$) yields an exact integral expression (eq. 2.50, see Appendix IV) which unfortunately cannot be reduced to closed form. On the other hand, playing the same trick of approximating the eigenvalues, we obtain the approximate pdf

$$p_{\mathbf{W}}(W|\mathbf{S}=S) \approx \frac{1}{2\sigma^2}\left(\frac{W}{\mu_{\text{eff}}S^2}\right)^{\frac{\mu_{\text{eff}}-1}{2}} \exp\left\{-\frac{W+\mu_{\text{eff}}S^2}{2\sigma^2}\right\} \mathrm{I}_{\mu_{\text{eff}}-1}\left(\frac{S\sqrt{\mu_{\text{eff}}W}}{\sigma^2}\right), \tag{2.19}$$

which again looks like (2.10) with μ_{eff} instead of μ.

Even though (2.18) and (2.19) look much simpler than (2.17) and (2.50) respectively, they still contain μ_{eff} as a parameter which in principle must be determined

from the diagonalization procedure outlined above. We now consider this problem from two viewpoints, one general, and one taking into account the *a priori* information from the diffraction-limited properties of the optical system.

Statistical moments

The following statements are straightforward to prove:

$$\mathrm{EV}\left\{\mathbf{I}(x)|\mathbf{S}=S\right\} = S^2 + 2\sigma^2, \tag{2.20}$$

$$\mathrm{Var}\left\{\mathbf{I}(x)|\mathbf{S}=S\right\} = 4\sigma^2\left(S^2+\sigma^2\right), \tag{2.21}$$

$$\mathrm{EV}\left\{\mathbf{W}|\mathbf{S}=S\right\} = S^2 + 2\sigma^2, \tag{2.22}$$

$$\mathrm{Var}\left\{\mathbf{W}|\mathbf{S}=S\right\} = \frac{4\sigma^2}{\mu^2}\sum_{n=1}^{\mu}\sum_{m=1}^{\mu}\left\{S^2\gamma_{11}(x_n-x_m)+\sigma^2\gamma_{11}^2(x_n-x_m)\right\}. \tag{2.23}$$

For dark pixels ($\mathbf{S}=0$) equations (2.20-2.23) hold by substituting $S=0$. Note that (2.23) was derived independently of the form of the pdf of $\mathbf{W}$, and therefore is in general exact even at the Riemann integral limit $\mu \to \infty$. In the limiting case where x_n, x_m are uncorrelated for $n \neq m$, (2.23) becomes simply

$$\mathrm{Var}\left\{\mathbf{W}|\mathbf{S}=S\right\} = \frac{4\sigma^2\left(S^2+\sigma^2\right)}{\mu}, \tag{2.24}$$

indicating a method to obtain μ_{eff} from image statistics: First we sample the image at a rate much higher than the correlation domain, so that $\mu >> \mu_{\mathrm{eff}}$. Equation (2.23) holds as is, whereas (2.24) holds with μ_{eff} in place of μ. By comparing the two equations, we obtain

$$\mu_{\mathrm{eff}} = \frac{\mathrm{Var}\left\{\mathbf{I}(x)|\mathbf{S}=S\right\}}{\mathrm{Var}\left\{\mathbf{W}|\mathbf{S}=S\right\}} = \frac{\mu^2\left(S^2+\sigma^2\right)}{\sum_{n=1}^{\mu}\sum_{m=1}^{\mu}\left\{S^2\gamma_{11}(x_n-x_m)+\sigma^2\gamma_{11}^2(x_n-x_m)\right\}}. \tag{2.25}$$

In order to be able to use relations (2.20-2.23) above, we must obtain all necessary quantities from experimental intensity measurements. Manipulating the above

relations we obtain:

$$S^2 = \mathbf{m}_{\mathbf{W},\mathbf{S}=S} - \mathbf{m}_{\mathbf{W},\mathbf{S}=0}, \tag{2.26}$$

$$\sigma^2 = \frac{1}{2}\mathbf{m}_{\mathbf{W},\mathbf{S}=0}, \tag{2.27}$$

$$\gamma_{11}(\xi|\mathbf{S}=0) = \sqrt{\frac{\gamma_{11}^{\mathbf{I}}(\xi|\mathbf{S}=0)}{\mathrm{Var}\,\{\mathbf{I}|\mathbf{S}=0\}} - 1}. \tag{2.28}$$

Again, these calculations are accurate, but not appropriate for practical situations, because they require many difficult measurements and a lot of computations. In diffraction-limited optical systems, it is easy to estimate μ_{eff} directly from the system apertures using diffraction theory, as we show in section 2.2.2.

Bit error rate

So far we have been concerned only with optical noise, which in a holographic memory is generated, for example, by light diffracted from neighboring pixels due to the finite modulation-transfer function of the optical system, by media scattering, laser speckle, etc. In addition, electrical noise [59, 60] is generated by photoelectric current fluctuations, shot noise, etc. Let v denote the total electrical signal on the detector. Similar to previous work [60], we model v as

$$v = v_{\mathrm{opt}} + v_{\mathrm{el}}, \tag{2.29}$$

where v_{opt} is the current generated by the optical signal (which contains only optical noise in the case of a dark – "OFF" – pixel, and mixture of signal plus noise in the case of a bright – "ON" – pixel) and v_{el} is the excess signal generated by electrical noise. We assume that the quantities v, v_{opt}, v_{el} are normalized to the amplitude variance σ^2 of the optical noise.

We now define the optical Signal-to-Noise Ratio $(\mathrm{SNR})_{\mathrm{opt}}$ as

$$(\mathrm{SNR})_{\mathrm{opt}} = \frac{S^2}{\sigma^2} = 2\left(\frac{\mathbf{m}_{\mathbf{W},\mathbf{S}=S}}{\mathbf{m}_{\mathbf{W},\mathbf{S}=0}} - 1\right). \tag{2.30}$$

A slightly different definition of SNR used often in optical systems is

$$(\mathrm{SNR})'_{\mathrm{opt}} = \frac{\mathbf{m}_{\mathbf{W},\mathbf{S}=S} - \mathbf{m}_{\mathbf{W},\mathbf{S}=0}}{\sqrt{\boldsymbol{\sigma}^2_{\mathbf{W},\mathbf{S}=S} + \boldsymbol{\sigma}^2_{\mathbf{W},\mathbf{S}=0}}}.$$

The two definitions are reconciled through the relation

$$(\mathrm{SNR})'_{\mathrm{opt}} = \frac{\sqrt{\mu_{\mathrm{eff}}}(\mathrm{SNR})^2_{\mathrm{opt}}}{2\sqrt{(\mathrm{SNR})^2_{\mathrm{opt}} + 2}}. \tag{2.31}$$

We will maintain definition (2.30) in later chapters because it conforms with the theory of Gaussian detection presented here.

The probability distributions of v_{opt} for the dark and bright pixels are given respectively by

$$p_{\mathrm{dark}}(v_{\mathrm{opt}}) = \left(\frac{\mu_{\mathrm{eff}}}{2}\right)^{\mu_{\mathrm{eff}}} \frac{v_{\mathrm{opt}}^{\mu_{\mathrm{eff}}-1}}{\Gamma(\mu_{\mathrm{eff}})} \exp\left\{-\frac{\mu_{\mathrm{eff}} v_{\mathrm{opt}}}{2}\right\}, \tag{2.32}$$

$$p_{\mathrm{bright}}(v_{\mathrm{opt}}) = \frac{1}{2}\left(\frac{v_{\mathrm{opt}}}{\mu_{\mathrm{eff}}(\mathrm{SNR})_{\mathrm{opt}}}\right)^{\frac{\mu_{\mathrm{eff}}-1}{2}} \exp\left\{-\frac{v_{\mathrm{opt}} + \mu_{\mathrm{eff}}(\mathrm{SNR})_{\mathrm{opt}}}{2}\right\} \times \mathrm{I}_{\mu_{\mathrm{eff}}-1}\left(\sqrt{\mu_{\mathrm{eff}}(\mathrm{SNR})_{\mathrm{opt}} v_{\mathrm{opt}}}\right). \tag{2.33}$$

The electrical SNR is defined as

$$(\mathrm{SNR})_{\mathrm{el}} = \frac{S^2}{\sigma^2_{\mathrm{el}}} = \frac{2\left(\mathbf{m}_{\mathbf{W},\mathbf{S}=S} - \mathbf{m}_{\mathbf{W},\mathbf{S}=0}\right)}{\sigma^2_{\mathrm{el}}}, \tag{2.34}$$

where σ^2_{el} is the variance of the electrical noise component. Because of the particular normalization of the electrical noise signal v_{el}, its approximate probability density function (independent of pixel value) is

$$p(v_{\mathrm{el}}) = \frac{(\mathrm{SNR})_{\mathrm{el}}}{\sqrt{2\pi}(\mathrm{SNR})_{\mathrm{opt}}} \exp\left\{-\frac{v^2_{\mathrm{el}}(\mathrm{SNR})^2_{\mathrm{el}}}{2(\mathrm{SNR})^2_{\mathrm{opt}}}\right\}. \tag{2.35}$$

The statistics of v are obtained by convolving $p(v_{\mathrm{el}})$ with $p_{\mathrm{dark}}(v_{\mathrm{opt}})$ or $p_{\mathrm{bright}}(v_{\mathrm{opt}})$ for dark or bright pixels respectively. The probability of error PE is obtained by

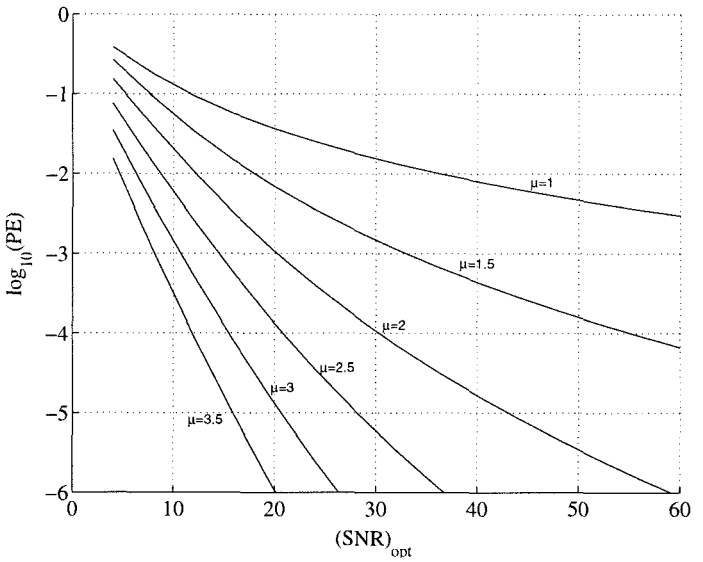


Figure 2.1: Probability of Error (PE) as function of optical signal to noise ratio $(\mathrm{SNR})_{\mathrm{opt}}$ for $(\mathrm{SNR})_{\mathrm{el}} = 10$ and different values of the noise coherence parameter μ.

standard Bayesian estimation on the resulting distributions and depends only on the parameters $(\mathrm{SNR})_{\mathrm{opt}}$, $(\mathrm{SNR})_{\mathrm{el}}$, and μ_{eff}. Some example plots of PE versus $(\mathrm{SNR})_{\mathrm{opt}}$ for $(\mathrm{SNR})_{\mathrm{el}} = 10$ and for various values of μ_{eff} are given in Fig. 2.1.

Generalizations

The above derivations were facilitated by the assumption that the bright and dark pixels follow the same statistics, and that the dark pixels have zero mean amplitude. In practical situations, the contrast ratio of spatial light modulators is finite, which means that the "dark" pixels are slightly biased in intensity. Also, the pixel statistics are usually non-stationary, bceause of beam non-uniformities and other nonidealities. These cases are more complicated, but they can still be treated in straightforward manner under the framework of this and subsequent sections; therefore, we will not consider them further in this thesis.

Appendix I: Derivation of the probability distribution for the integrated intensity of a bright pixel

We will use the method of characteristic functions to derive the probability density function of the random variable $\mathbf{W}$ given by (2.5) where the components $\mathbf{I}(x_n)$ are

independent random variables with $\mathbf{S} = S$. Let $\mathbf{I}$ denote any one of the intensity variables. The characteristic function is defined as

$$\Psi_{\mathbf{I}}(t) = \mathrm{EV}\left\{\exp\left\{i\mathbf{I}t\right\}\right\}. \tag{2.36}$$

It is convenient to define the auxiliary variables $\boldsymbol{\rho}$, $\boldsymbol{\theta}$ according to

$$\begin{aligned} \mathbf{N}_{\mathrm{r}} + i\mathbf{N}_{\mathrm{i}} &= \boldsymbol{\rho}\mathrm{e}^{i\boldsymbol{\theta}}, \quad \text{where} \\ \boldsymbol{\rho}^2 &= \mathbf{N}_{\mathrm{r}}^2 + \mathbf{N}_{\mathrm{i}}^2, \quad \text{and} \\ \boldsymbol{\theta} &= \arctan\frac{\mathbf{N}_{\mathrm{i}}}{\mathbf{N}_{\mathrm{r}}}. \end{aligned}$$

In the $(\boldsymbol{\rho}, \boldsymbol{\theta})$ space, eq. 2.36 becomes

$$\Psi_{\mathbf{I}}(t) = \exp\left\{iS^2t\right\}\mathrm{EV}\left\{\exp\left\{i\left(\boldsymbol{\rho}^2 + 2S\boldsymbol{\rho}\cos\boldsymbol{\theta}\right)t\right\}\right\}. \tag{2.37}$$

Under the current assumptions, $\boldsymbol{\rho}$ is Rayleigh-distributed in $[0, \infty)$ and $\boldsymbol{\theta}$ is uniform in $[-\pi, \pi)$. Therefore $\Psi_{\mathbf{I}}(t)$ can be written explicitly in integral form as

$$\Psi_{\mathbf{I}}(t) = \mathrm{e}^{iS^2t}\int_0^\infty \mathrm{d}\rho \int_{-\pi}^{\pi} \mathrm{d}\theta \frac{\rho}{2\pi\sigma^2}\mathrm{e}^{-\rho^2/2\sigma^2}\mathrm{e}^{i\rho(\rho+i2S\cos\theta)t}. \tag{2.38}$$

Using one of the integral definitions of Bessel functions [61],

$$\mathrm{J}_n(z) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-ni\phi+iz\sin\phi}\mathrm{d}\phi,$$

we reduce (2.38) to

$$\Psi_{\mathbf{I}}(t) = \mathrm{e}^{iS^2t}\int_0^\infty \frac{\rho}{\sigma^2}\exp\left\{-\left(\frac{1}{2\sigma^2} - it\right)\rho^2\right\}\mathrm{J}_0\left(2S\rho\, t\right)\mathrm{d}\rho. \tag{2.39}$$

From standard mathematical tables ([62] formula 6.631-4) and after some manipula-

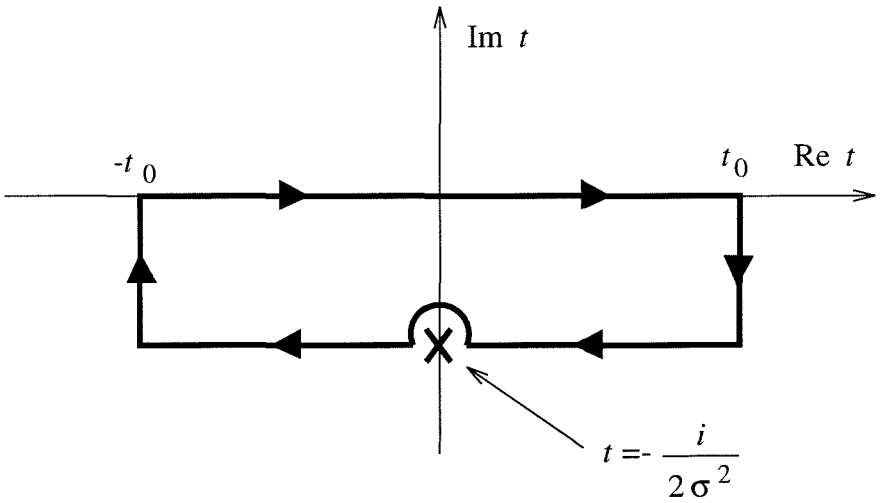


Figure 2.2: Path integral used for the evaluation of (2.42).

tion we obtain the characteristic function in closed form

$$\Psi_{\mathbf{I}}(t) = \frac{1}{1 - 2i\sigma^2 t} \exp\left\{\frac{iS^2 t}{1 - 2i\sigma^2 t}\right\}. \tag{2.40}$$

Since we are assuming here that the μ components in (2.5) are independent, the characteristic function for $\mathbf{W}$ is simply

$$\Psi_{\mathbf{W}}(t) = (\Psi_{\mathbf{I}}(t))^{\mu} = \frac{1}{(1 - 2i\sigma^2 t)^{\mu}} \exp\left\{\frac{i\mu S^2 t}{1 - 2i\sigma^2 t}\right\}. \tag{2.41}$$

We now must calculate the inverse Fourier transform of (2.41) from the integral

$$\begin{aligned} p_{\mathbf{W}}(W) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_{\mathbf{W}}(t) \mathrm{e}^{-iWt} \mathrm{d}t \\ &= \int_{-\infty}^{\infty} \frac{1}{(1 - 2i\sigma^2 t)^{\mu}} \exp\left\{i\left(\frac{\mu S^2}{1 - 2i\sigma^2 t} - W\right) t\right\} \mathrm{d}t. \end{aligned} \tag{2.42}$$

To calculate this integral we consider the loop integral along the closed path shown in Fig. 2.2 with $t_0 \to \infty$. This integral is zero, because the function is analytic in the interior of the path (note that the pole $t = -i/2\sigma^2$ has been excluded). We recognize the part of the path that lies on the real axis as the path of integration appearing

in (2.42). Consider the integral along the vertical part of the path at t_0. If we let $t = t_0 + i\zeta$ along the path, the integral after some manipulation can be written as

$$\int_0^{-\frac{1}{2\sigma^2}} \frac{1}{(1+2\sigma^2\zeta - 2i\sigma^2 t_0)^M} \exp\left\{-\mu S^2 \frac{2\sigma^2 t_0^2 + \zeta(1+2\sigma^2\zeta) - it_0}{(1+2\sigma^2\zeta)^2 + 4\sigma^4 t_0^2} - iWt\right\} \mathrm{d}\zeta.$$

This clearly goes to zero as $t_0 \to \infty$, because the non-oscillatory part of the exponential tends to one and is therefore dominated by the preceding fraction. For the same reason the integral along the left-hand side vertical path ($t = -t_0 + i\zeta$) goes to zero. The integral along the remainder of the path is

$$\exp\left\{-\frac{\mu S^2 + W}{2\sigma^2}\right\} \oint_+ \frac{1}{(-2i\sigma^2 u)^\mu} \exp\left\{i\left(\frac{\mu S^2}{4\sigma^4 u} - Wu\right)\right\} \mathrm{d}u,$$

where $\oint_+$ denotes integration along the path that excludes the pole $1 - 2i\sigma^2 t = 0$ in the positive (counter-clockwise) direction, as shown in Fig. 2.2. We can re-write the integral as follows:

$$\exp\left\{-\frac{\mu S^2 + W}{2\sigma^2}\right\} \oint_+ \frac{1}{(-2i\sigma^2 u)^\mu} \exp\left\{i\frac{S\sqrt{\mu W}}{2\sigma^2}\left(\frac{\sqrt{\mu}S}{2\sigma^2\sqrt{W}u} - \frac{2\sigma^2\sqrt{W}u}{\sqrt{\mu}S}\right)\right\} \mathrm{d}u =$$

$$= \frac{i^\mu}{2\sigma^2}\left(\frac{W}{\mu S^2}\right)^{\frac{\mu-1}{2}} \exp\left\{-\frac{\mu S^2 + w}{2\sigma^2}\right\} \oint_+ u^{-\mu} \exp\left\{-i\frac{S\sqrt{\mu W}}{2\sigma^2}\left(u - \frac{1}{u}\right)\right\} \mathrm{d}u.$$

The last integrand happens to be the generating function of a Laurent series with Bessel functions J_ν as coefficients ([63], page 355):

$$\mathrm{J}_\nu(z) = \frac{1}{2\pi i} \oint_+ u^{-(\nu+1)} \exp\left\{\frac{z}{2}\left(u - \frac{1}{u}\right)\right\} \mathrm{d}u.$$

Substituting, and using the relation $\mathrm{I}_\nu(z) = i^{-\nu}\mathrm{J}_\nu(iz)$ we obtain (2.10), i.e.,

$$p_{\mathbf{W}}(W) = \frac{1}{2\sigma^2}\left(\frac{W}{\mu S^2}\right)^{\frac{\mu-1}{2}} \exp\left\{-\frac{W + \mu S^2}{2\sigma^2}\right\} \mathrm{I}_{\mu-1}\left(\frac{S\sqrt{\mu W}}{\sigma^2}\right).$$

Appendix II: Derivation of the exact probability density function for the integrated intensity of a dark pixel

We begin by defining the intensity process in the decoupled domain

$$\tilde{\mathbf{I}}(x_j) = \tilde{\mathbf{N}}_{\mathrm{r}}^2(x_j) + \tilde{\mathbf{N}}_{\mathrm{i}}^2(x_j), \qquad j = 1, \ldots, \mu. \tag{2.43}$$

The pdf of $\tilde{\mathbf{I}}(x_j)$ is expressed in form similar to (2.7) as follows:

$$p_{\tilde{\mathbf{I}}}(\tilde{I}) = \frac{1}{2\sigma^2} \exp\left\{-\frac{\tilde{I}}{2\sigma^2}\right\}. \tag{2.44}$$

The integrated intensity $\mathbf{W}$ can be expressed in terms of the noise vectors as

$$\begin{aligned} \mu\mathbf{W} &= \mathcal{N}_{\mathrm{r}}^T\mathcal{N}_{\mathrm{r}} + \mathcal{N}_{\mathrm{i}}^T\mathcal{N}_{\mathrm{i}} \\ &= \tilde{\mathcal{N}}_{\mathrm{r}}^T Q^T Q\tilde{\mathcal{N}}_{\mathrm{r}} + \tilde{\mathcal{N}}_{\mathrm{i}}^T Q^T Q\tilde{\mathcal{N}}_{\mathrm{i}} \\ &= \tilde{\mathcal{N}}_{\mathrm{r}}^T\tilde{\mathcal{N}}_{\mathrm{r}} + \tilde{\mathcal{N}}_{\mathrm{i}}^T\tilde{\mathcal{N}}_{\mathrm{i}} \\ &= \sum_{m=1}^{\mu} \tilde{\mathbf{I}}(x_j). \end{aligned} \tag{2.45}$$

In the third step we used the fact that Q is unitary. Since the $\tilde{\mathbf{I}}(x_j)$'s are independent, the convolution method can be applied to calculate the pdf of their sum (2.45). The characteristic function of (2.44) is

$$\Psi_{\tilde{I}(x_j)}(t) = \frac{1}{1 - 2i\lambda_j^2 t}. \tag{2.46}$$

The characteristic function of $\mu\mathbf{W}$ is expressed as a product

$$\begin{aligned} \Psi_{\mu\mathbf{W}}(t) &= \prod_{m=1}^{\mu} \frac{1}{1 - 2i\lambda_m^2 t} \\ &= \sum_{m=0}^{\mu} \frac{1}{(1 - 2i\lambda_m^2 t)\prod_{m' \neq m}\left(1 - \frac{\lambda_{m'}^2}{\lambda_m^2}\right)}. \end{aligned} \tag{2.47}$$

Direct Fourier inversion of (2.47) yields (2.17).

Appendix III: Distribution tails for dark pixels

In this appendix we will argue that the approximations (2.18), (2.19) become

arbitrarily accurate at the tails of the distributions, i.e., as $W \to \infty$ for (2.18) and as $W \to 0$ for (2.19). We begin with the dark pixels, whose exact distribution is given by (2.17). Suppose that the eigenvalues are unevenly distributed with a few ($\approx \mu_{\text{eff}}$) having values close to $2\sigma^2$ and the rest with values close to zero. From (2.17) it follows that the contribution of the μ-th component in the decoupled space can be made smaller than an arbitrary number, say δ, if

$$W > -2\lambda_m^2 \log \left[2\delta \prod_{m' \neq m} \left| \lambda_m^2 - \lambda_{m'}^2 \right| \right].$$

As λ_m decreases, the lower bound for W decreases as well; therefore, the contributions of near-zero eigenvalues become insignificant towards the right-hand tail of the distribution. This claim is also supported by the numerical results presented in [56].

Consider now the integral expression (2.50) for the pdf of the integrated intensity of a bright pixel. Each exponential contributes most strongly for t being in the region of stationarity, i.e.,

$$\frac{\partial}{\partial t} \left\{ \frac{i\tilde{S}_m^2 t}{1 - 2\lambda_m^2 t} + \frac{iWt}{\mu} \right\} = 0$$

$$\to t = -\frac{i}{2\lambda_m^2} - \frac{\tilde{S}_m^2}{2W\lambda_m^2}$$

We can see that as W and λ_m decrease, the stationary points move away from the real axis, while the real parts of the exponents become small. Therefore, the small eigenvalues contribute negligibly as $w \to 0$.

Appendix IV: Derivation of the exact probability density function for the integrated intensity of a bright pixel

We define the signal and electric field vectors $\mathcal{S}$ and $\mathcal{E}$ respectively as follows:

$$\mathcal{S} = \underbrace{(S, \ldots, S)}_{\mu \text{ elements}}$$

$$\mathcal{E} = \mathcal{E}_{\text{r}} + \mathcal{E}_{\text{i}} = \mathcal{S} + \mathcal{N}_{\text{r}} + i\mathcal{N}_{\text{i}}$$

The pdf for the field vector is Gaussian

$$p_{\mathcal{E}_r\mathcal{E}_i}(\overline{e}_r, \overline{e}_i) = \frac{1}{(2\pi)^\mu |\Sigma|} \exp\left\{-\frac{1}{2}(\overline{e}_r - \mathcal{S})^T \Sigma^{-1} (\overline{e}_r - \mathcal{S}) - \frac{1}{2}\overline{e}_i^T \Sigma^{-1} \overline{e}_i\right\}. \tag{2.48}$$

After diagonalizing Σ as in the case of dark pixels (eq. 2.13) and defining $\tilde{\mathcal{S}} = Q\mathcal{S}$, $\tilde{\mathcal{E}} = Q\mathcal{E}$, we obtain the decoupled pdf

$$p_{\tilde{\mathcal{E}}_r\tilde{\mathcal{E}}_i}(\overline{e}_r, \overline{e}_i) = \frac{1}{(2\pi)^\mu |\Lambda|} \exp\left\{-\frac{1}{2}\left(\overline{e}_r - \tilde{\mathcal{S}}\right)^T \Sigma^{-1} \left(\overline{e}_r - \tilde{\mathcal{S}}\right) - \frac{1}{2}\overline{e}_i^T \Lambda^{-1} \overline{e}_i\right\}. \tag{2.49}$$

We then use the property $\mathcal{E}^T\mathcal{E} = \tilde{\mathcal{E}}^T\tilde{\mathcal{E}}$ (see Appendix 2.2.1) to obtain $\mathbf{W}$ as a sum of independent intensity variables $\tilde{\mathbf{I}}(x_j)$. The characteristic function for those is similar to (2.8) (but with λ_j instead of σ) and the characteristic function for $\mu\mathbf{W}$ is obtained as the product of the μ individual pdf's. Finally, the pdf is obtained in integral form as

$$p_{\mathbf{W}}(w) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \prod_{m=1}^{\mu} \left(\frac{1}{1-2\lambda_m^2 t} \exp\left\{\frac{i\tilde{S}_m^2 t}{1-2\lambda_m^2 t} + \frac{iwt}{\mu}\right\}\right) \mathrm{d}t. \tag{2.50}$$

2.2.2 Diffraction-limited optical noise

In the previous section we analyzed the spatial optical noise in a page detected by an optical system, and saw the effect of the band-limited nature of the noise on the probability of error, through the concept of the "effective degrees of freedom," μ_{eff}. Here we give an example of the calculation of μ_{eff} in a simple optical system, and justify the approximations leading to (2.18) and (2.19). A similar calculation for time-domain integration is given in [56].

Consider an ideal, noiseless monochromatic point source located distance d behind a circular aperture of radius R, as in Figure 2.3. The radiation is affected by (spatial) white noise in the form of a thin transparency located infinitesimally close to the point source. We are interested in the statistical properties of the noise immediately

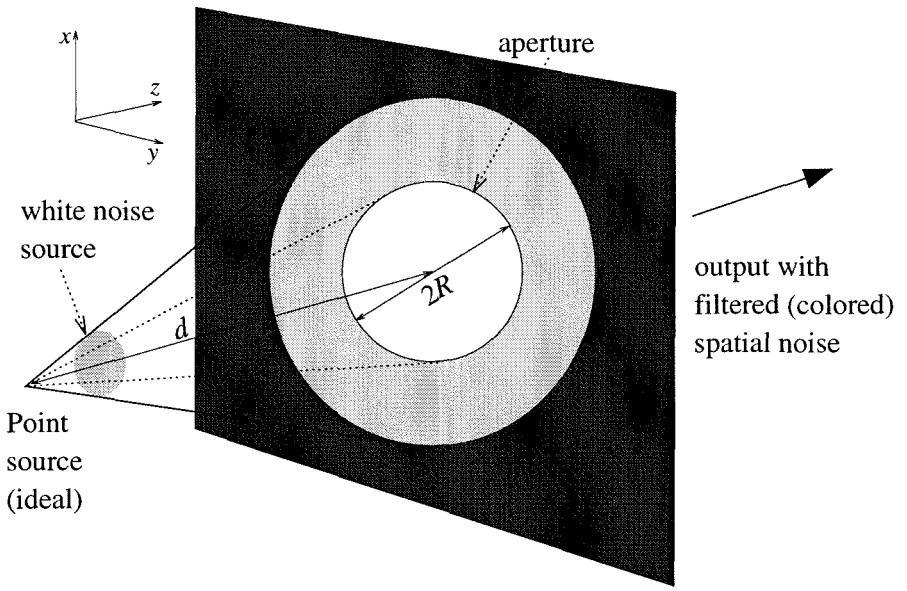


Figure 2.3: Geometry for the calculation of the effect of an aperture on spatial white noise.

past the aperture, and in particular the size of the "correlation domains," which we later use to determine μ_{eff}.

Each infinitesimal point source on the noise transparency is spatially filtered by the aperture, producing a diffraction pattern. An immediate result of this statement is that it is impossible to distinguish the individual noise sources contributing noise to an area equal to the minimum resolvable spot (MRS) of the aperture. In other words, despite the fact that the noise sources were originally uncorrelated (since the noise source was assumed spatially white), still the noise contributions in distinct points separated by less than the MRS are correlated. We may therefore claim that the number of effective degrees of freedom μ_{eff} equals the amount of MRS's fitting inside one pixel, since this is the amount of effectively uncorrelated random variables contributing to the integrated detector signal. Our objective in the remainder of this section is to quantify these statements.

For simplicity, we will still perform the analysis for a one-dimensional noise pattern beyond an aperture of size $2R$, and correct for the circular aperture afterwards. The

autocorrelation function immediately past the aperture is, by definition,

$$\gamma_{11}(x_1, x_2; d^+) = \mathrm{EV}\left\{\mathbf{E}(x_1; d^+)\mathbf{E}^*(x_2; d^+)\right\}, \tag{2.51}$$

where in the notation we added the distance d from the point source, and the superscript "+" denotes "immediately after." The electric field is expressed explicitly using Fresnel diffraction theory as

$$\mathbf{E}(x; d^+) = \int_{-\infty}^{+\infty} \mathrm{rect}\left(\frac{x'}{2R}\right) \mathrm{e}^{i\pi\frac{(x-x')^2}{\lambda d}} \mathrm{d}x', \tag{2.52}$$

where some constant factors were omitted for simplicity without affecting the calculation. Substituting into (2.51) we obtain

$$\gamma_{11}(x_1, x_2; d^+) = \mathrm{EV}\left\{\iint_{-\infty}^{+\infty} \mathrm{rect}\left(\frac{x_1'}{2R}\right) \mathrm{rect}\left(\frac{x_2'}{2R}\right) \mathbf{E}(x_1'; 0^+)\mathbf{E}^*(x_2'; 0^+) \right.$$
$$\left. \exp\left\{i\pi\frac{(x_1 - x_1')^2 - (x_2 - x_2')^2}{\lambda d}\right\} \mathrm{d}x_1'\mathrm{d}x_2'\right\}. \tag{2.53}$$

By the definition of a white noise source, we obtain that at the origin the noise correlation is

$$\gamma_{11}(x_1', x_2'; 0^+) = \mathrm{EV}\left\{\mathbf{E}(x_1'; 0^+)\mathbf{E}^*(x_2'; 0^+)\right\} = \delta(x_1' - x_2'). \tag{2.54}$$

After exchanging the orders of the expectation value and integration operators in (2.53), one of the integrations is performed trivially because of the δ-function, and we obtain

$$\gamma_{11}(x_1, x_2; d^+) = \mathrm{e}^{i\pi\frac{x_1^2 - x_2^2}{\lambda d}} \int_{-\infty}^{+\infty} \mathrm{rect}\left(\frac{x'}{2R}\right) \mathrm{e}^{-i2\pi\frac{(x_1 - x_2)x'}{\lambda d}} \mathrm{d}x' =$$
$$= \mathrm{e}^{i2\pi\frac{\xi\bar{\xi}}{\lambda d}} \mathrm{sinc}\left(\frac{2\xi R}{\lambda d}\right), \qquad \xi \equiv x_1 - x_2, \qquad \bar{\xi} \equiv \frac{x_1 + x_2}{2}. \tag{2.55}$$

The "correlation domain" for our simple one-dimensional case is the size of the first

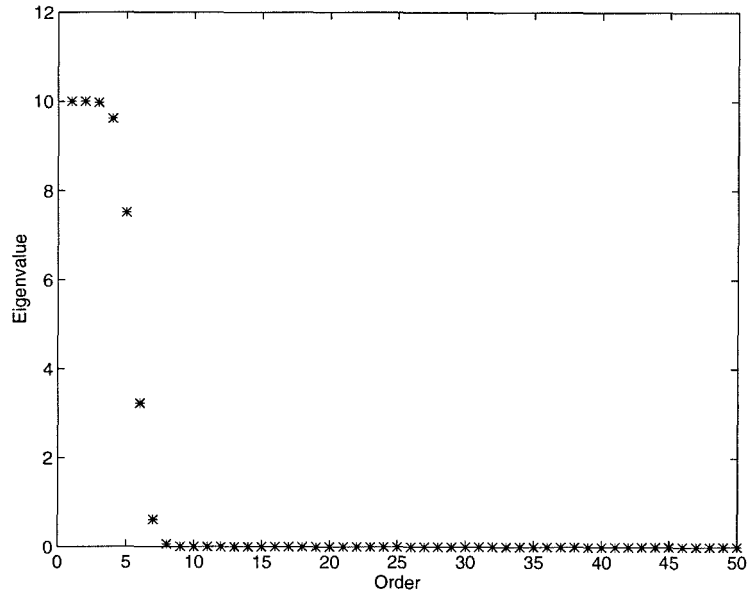


Figure 2.4: Distribution of eigenvalue magnitudes for an autocorrelation matrix of a diffraction limited system with input spatial white noise and spatial detector integration at the output.

diffraction lobe, i.e., $\lambda d/2R$. For the circular aperture of Fig. 2.3, repeating the above calculation in two dimensions yields the diameter of the correlation domain

$$(\Delta\xi)_c = 1.22\frac{\lambda d}{R}. \tag{2.56}$$

Note that both formulas are identical to the MRS's of the corresponding optical systems.

Now consider an optical system affected by spatial noise with autocorrelation given by (2.55). We form the autocorrelation matrix Σ according to (2.12), and observe that if the pixel size is $b = n\,(\Delta\xi)_c$, then approximately n eigenvalues of Σ are approximately equal to 1 and the rest are approximately equal to zero. An example is given in Figure 2.4, where the (one-dimensional) pixel size was chosen to be five times the correlation domain, and $|\gamma_{11}(\xi)|$ was sampled at $\mu = 50$ individual points. This picture justifies the approximations that led to (2.18), (2.19), and also

the estimate

$$\mu_{\text{eff}} \approx \left(\frac{b}{(\Delta\xi)_c} \right)^2 . \tag{2.57}$$

2.3 Noise and surface density

From the analysis in the previous section it follows that making the pixel size b large helps reduce the noise, because then the effective degrees of freedom μ_{eff} grow thereby improving the effect of detector integration. At the same time, though, increasing b reduces the effective space-bandwidth product of the optical system. Therefore a trade-off emerges: one can design a high density system tolerating some extra noise, and use error correction to undo noise effects; however, codes produce overhead to the stored information thereby also decreasing the density. Error correction codes for page-oriented memories is currently an active research topic [64, 65, 66, 67]. Here we will confine ourselves to the calculation of the upper limit for the useful information that can be stored in a holographic memory[3], which is easily calculated using Shannon's coding theorem.

We will treat the holographic memory as a binary asymmetric channel, as shown in Figure 2.5. Each piece of information is stored as either 0, with probability ω, or 1, with probability $1-\omega$. At detection, there are finite probabilities ε_{01} that 0 turns into a 1, and ε_{10} that a 1 turns into a 0. Let $H(\text{detect})$ denote the entropy of the detected bits, and $H(\text{detect} \mid \text{store})$ the a posteriori entropy of the detected bits given the stored bits (in other words the uncertainty about the output given the input). The mutual information of the input and output sets (i.e., the amount of information that we can infer from the detected data about the stored data and vice versa) is defined as

$$M(\text{store}, \text{detect}; \omega) = H(\text{detect}) - H(\text{detect} \mid \text{store}). \tag{2.58}$$

[3] i.e., the amount of information remaining available after the error correction overhead is subtracted from the raw density.

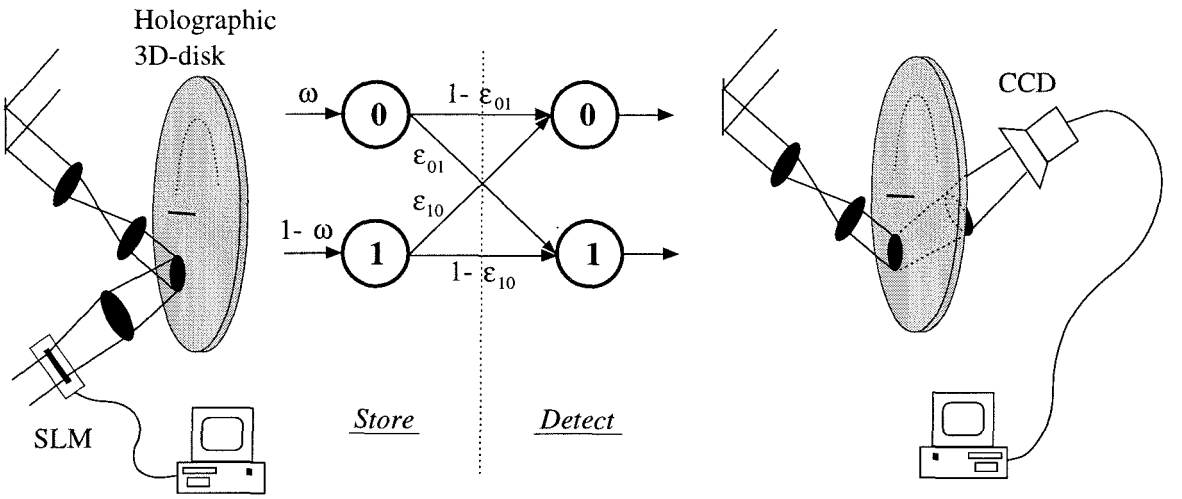


Figure 2.5: A holographic memory viewed as an information channel.

The mutual information is maximized if ω is selected as

$$\omega_{\text{opt}} = \frac{1-(\nu+1)\varepsilon_{10}}{(\nu+1)(1-\varepsilon_{01}-\varepsilon_{10})}, \tag{2.59}$$

where $\log_2 \nu = \dfrac{\mathcal{H}(\varepsilon_{10}) - \mathcal{H}(\varepsilon_{01})}{1-\varepsilon_{01}-\varepsilon_{10}}$, $\mathcal{H}(w) = -w\log_2 w - (1-w)\log_2(1-w)$.

This relation simply says that it is better to select a coding scheme that favors the symbol (0 or 1) that is less likely to flip between storage and detection in the asymmetric channel. If $\varepsilon_{01} = \varepsilon_{10}$, then we obtain the intuitive result $\omega = 1/2$, which is common in practical coding schemes. Returning to the optimal result (2.59), Shannon's theorem states that a coding scheme exists such that a maximum amount of data equal to $M(\text{store, detect}; \omega_{\text{opt}})$ with arbitrarily small probability of error. Thus $M(\text{store, detect}; \omega_{\text{opt}})$ is an upper bound on the amount of useful information (as opposed to error correction overhead) that one may store in the memory channel of Fig. 2.5.

To understand the trade-off between noise and surface density, let us consider an extreme case of noisy data, e.g., $S/\sigma = 1.9$, with optical noise only. The parameters μ, ε_{01}, ε_{10} are calculated using the theory (equations 2.57, 2.18, 2.19) of section 2.2.

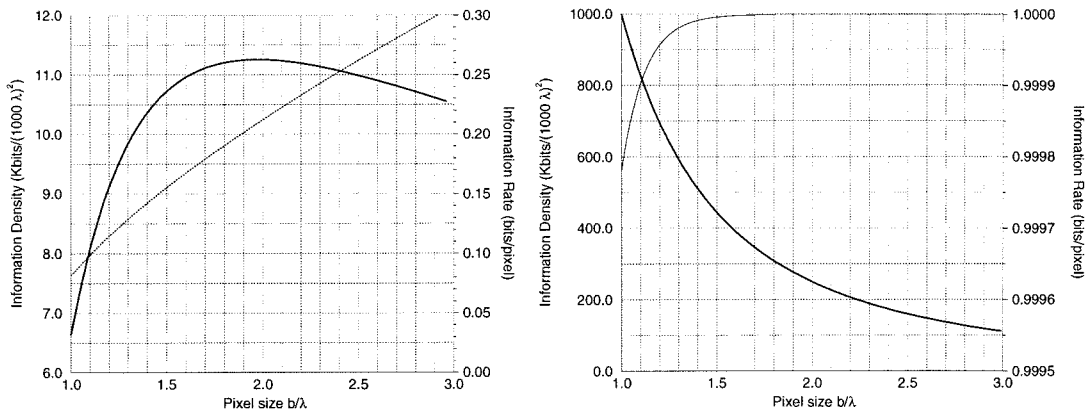


Figure 2.6: Examples of tradeoff between noise and storage density in page-oriented optical memories. The plots show *the upper limits* in the amount of information bits that can be stored in a page of fixed size equal to 1000 wavelengths, and the ratio of useful information versus error-correction overhead, versus the pixel size b normalized to the wavelength λ. It is assumed that the bandwidth of the optical system is enough to avoid vigneting or filtering effects. The optical SNR is (a) 1.9, (b) 10.

The effective capacity versus normalized pixel size (relative to the wavelength) is shown in Fig. 2.6a. We observe that, for small pixel sizes, detector integration is not sufficient to cancel the noise, hence the error correction overhead limits the density severely. On the other hand, if the pixel size is large, the overhead is small but the density is reduced by the fact that not enough pixels fit in one page. An optimum is found when $b \approx 2\lambda$.

On the other hand, if the noise level is sufficiently low, e.g., $S/\sigma = 10$ (Fig. 2.6b), then the error correcting overhead is never significant. In that case, one is better off using pixel size as small as possible[4].

The above conclusions are true when the memory contains a single page per location. In Chapter 5 and section 6.2 we will see that additional considerations must be taken into account in holographic memories, where several overlapping pages are multiplexed, because then the density and noise are affected by the pixel size in a more complicated manner. The information metric developed here will then help in clearing up the associated trade-offs.

[4]i.e., as small as allowed by the technology

Chapter 3 Volume holography with non-planar reference waves

3.1 Fundamentals of volume holography

Volume holograms are stored as a result of interference between two mutually coherent light beams, the signal and the reference (Fig. 3.1). The signal carries the information, typically in the form of amplitude modulation imprinted on the wavefront. In the simplest case, the reference is a plane wave (in this chapter we discuss two cases of non-planar references). In a thick medium one needs to reproduce the reference used for recording as accurately as possible in order to get diffraction from the hologram. If instead the plane-wave reference deviates in angle or wavelength, then diffraction contributions from different parts of the hologram become phase mismatched causing the diffraction efficiency to drop (Bragg mismatch). The amount by which the angle or wavelength need to change before the reconstructed power drops to zero is called Bragg selectivity and depends on the geometry and the thickness of the material.

As an example[1], consider the transmission geometry of Fig. 3.1, a very common setup for holographic storage. The plane-wave reference $R(\boldsymbol{r})$ is incident at angle θ_R, the signal $S(\boldsymbol{r})$ at θ_S, and they are both at wavelength λ_R. The hologram is a perturbation of the refractive index ϵ by an amount $\Delta\epsilon$ proportional to the intensity, i.e., the interference pattern that results from the coherent superposition of the reference and signal. In the simple example of Fig. 3.1, we have:

$$R(\boldsymbol{r}) = \exp\{i\boldsymbol{k}_R \cdot \boldsymbol{r}\}, \quad \boldsymbol{k}_R = \frac{2\pi}{\lambda}(-\sin\theta_R, 0, \cos\theta_R); \tag{3.1}$$

$$S(\boldsymbol{r}) = \exp\{i\boldsymbol{k}_S \cdot \boldsymbol{r}\}, \quad \boldsymbol{k}_S = \frac{2\pi}{\lambda}(\sin\theta_S, 0, \cos\theta_S); \tag{3.2}$$

[1]The theory of volume holography presented in this section is based on the (unpublished) class notes and homeworks of APh/EE 133 (Optical Computing), by D. Psaltis, Caltech 1994.

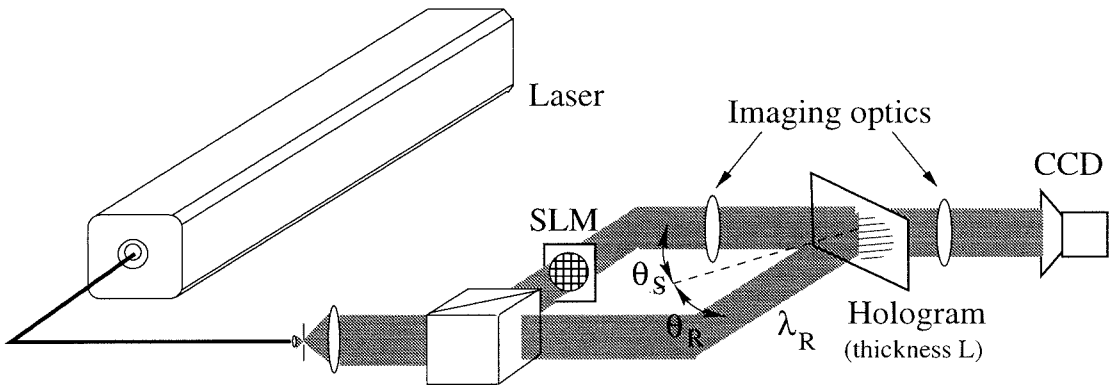


Figure 3.1: Volume holographic memory in the transmission geometry.

$$\begin{aligned}\Delta\epsilon(\boldsymbol{r}) =& \epsilon_1 \left|R(\boldsymbol{r}) + S(\boldsymbol{r})\right|^2 = 2\epsilon_1 \left\{1 + \cos\left(\boldsymbol{K}_g \cdot \boldsymbol{r}\right)\right\}, \quad \boldsymbol{K}_g = \boldsymbol{k}_S - \boldsymbol{k}_R; \\ \Delta\epsilon(\boldsymbol{r}) =& \ 2\epsilon_1 \left\{1 + \cos\left[\left(\sin\theta_S + \sin\theta_R\right)\frac{2\pi x}{\lambda} + \left(\cos\theta_S - \cos\theta_R\right)\frac{2\pi z}{\lambda}\right]\right\}.\end{aligned} \tag{3.3}$$

Here ϵ_1 denotes the strength of the hologram, $\boldsymbol{r} = (x, y, z)$ is the position vector, and $\boldsymbol{K}_g$ is the grating vector. For later convenience, we break $\Delta\epsilon(\boldsymbol{r})$ into three terms as follows:

$$\Delta\epsilon(\boldsymbol{r}) = \Delta\epsilon_0(\boldsymbol{r}) + \Delta\epsilon_+(\boldsymbol{r}) + \Delta\epsilon_-(\boldsymbol{r}),$$

$$\Delta\epsilon_0(\boldsymbol{r}) = 2\epsilon_1, \quad \Delta\epsilon_+(\boldsymbol{r}) = \epsilon_1 \exp\left\{i\boldsymbol{K}_g \cdot \boldsymbol{r}\right\}, \quad \Delta\epsilon_-(\boldsymbol{r}) = \epsilon_1 \exp\left\{-i\boldsymbol{K}_g \cdot \boldsymbol{r}\right\}.$$

The reconstruction of the recorded hologram is posed as the problem of diffraction from the pattern $\Delta\epsilon(\boldsymbol{r})$ upon illumination by a reconstructing reference beam $\mathcal{E}_{R'}(\boldsymbol{r})$ (possibly different than R). This problem admits analytical solution under the Born and paraxial approximations [68, 69, 70]. The first-order distortion introduced to the hologram by non-paraxial terms is calculated in Appendix I. The paraxial result for the diffracted field $\mathcal{E}_d$ is:

$$\mathcal{E}_d(\boldsymbol{r}) = \iint_{-\infty}^{+\infty} \mathbf{A}(\boldsymbol{k}_{R'}; \boldsymbol{k}_d) \exp\left\{i\boldsymbol{k}_d \cdot \boldsymbol{r}\right\} \mathrm{d}k_{dx}\mathrm{d}k_{dy}, \tag{3.4}$$

$$\text{where} \quad \mathbf{A}(\boldsymbol{k}_{R'};\boldsymbol{k}_d) = \boldsymbol{a}\iiint_{\mathcal{V}} \Delta\epsilon_+(\boldsymbol{r}')\exp\left\{i\left(\boldsymbol{k}_{R'}-\boldsymbol{k}_d\right)\cdot\boldsymbol{r}'\right\}\mathrm{d}^3\boldsymbol{r}', \tag{3.5}$$

where the constant vector $\boldsymbol{a}$ is given by

$$\boldsymbol{a}\left(\boldsymbol{K}_g\right) = \frac{\Delta\epsilon_0}{2\,i\,k_{dz}}\left\{\frac{\boldsymbol{K}_g\cdot\hat{\mathbf{e}}_\mathbf{i}}{\epsilon_0}\left(\boldsymbol{k}_{R'}+\boldsymbol{K}_g\right)-\omega^2\mu_0\epsilon\hat{\mathbf{e}}_\mathbf{i}\right\} \tag{3.6}$$

($\hat{\mathbf{e}}_\mathbf{i}$ is the polarization vector of the incident electric field). In (3.4) the integral is over the space of possible difracted wave-vectors

$$\boldsymbol{k}_d = \left(k_{dx}, k_{dy}, \sqrt{\left(\frac{2\pi}{\lambda_{R'}}\right)^2 - k_{dx}^2 - k_{dy}^2}\right).$$

In contrast to infinitely thin holograms (Raman-Nath diffraction regime), only the term $\Delta\epsilon_+$ contributes to the diffracted field. The other two terms $\Delta\epsilon_0$, $\Delta\epsilon_-$ can be shown to have negligible contribution. The constant vector $\boldsymbol{a}$ of (3.4) and (3.5) is a result of the vectorial nature of diffraction (details on the calculation of $\boldsymbol{a}$ are given in [69]). $\mathcal{V}$ is the three-dimensional extent of the holographic material. The factor $\mathbf{A}(\boldsymbol{k}_{R'};\boldsymbol{k}_d)$ is interpreted as the 3–D Fourier transform of the hologram calculated at spatial frequency $\boldsymbol{k}_{R'}-\boldsymbol{k}_d$.

This integral formulation is convenient for calculating diffraction from holograms recorded with a reference beam that is a plane wave or a discrete superposition of plane waves (see, e.g., section 3.2). For other wavefronts (e.g., spherical), the calculation is quite formidable. However, using standard properties of Fourier integrals, (3.4) and (3.5) can be readily re-expressed in convolution form:

$$\mathcal{E}_d(\boldsymbol{r}) = \iiint_{\mathcal{V}} \mathcal{E}_{R'}(\boldsymbol{r}')\Delta\epsilon_+(\boldsymbol{r}')\mathcal{G}(\boldsymbol{r}';\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}'. \tag{3.7}$$

$$\mathcal{G}(\boldsymbol{r}';\boldsymbol{r}) = \frac{1}{i\lambda_{R'}\left|\boldsymbol{r}-\boldsymbol{r}'\right|}\exp\left\{i\,k_{R'}\,\left|\boldsymbol{r}-\boldsymbol{r}'\right|\right\}, \qquad \boldsymbol{r}\neq\boldsymbol{r}' \tag{3.8}$$

is the scalar form of Green's function for free space. Eq. (3.7) has an interesting interpretation [70]: it says that each infinitesimal region inside the hologram acts as a point-source; the diffracted field is obtained as coherent superposition of all the

infinitesimal point-sources comprising the hologram. We will use this particular form later (sections 3.3.1–Appendix, 4.1) for the calculation of diffraction from volume holograms recorded with a spherical wave reference.

Returning to the simple example of Fig. 3.1, suppose that the holographic material has thickness L in the z-direction and is infinite in the transverse directions x, y, and that the incident field is $R'(\boldsymbol{r})$ with wave-vector $\boldsymbol{k}_{R'}$. Applying (3.4-3.5) we find that the diffracted wave-vector $\boldsymbol{k}_d$ is determined by the following conditions:

$$\boldsymbol{k}_d \times \hat{\mathbf{z}} = (\boldsymbol{k}_{R'} + \boldsymbol{K}_g) \times \hat{\mathbf{z}}, \tag{3.9}$$

$$|\boldsymbol{k}_d| = k_{R'} = 2\pi/\lambda_{R'}, \tag{3.10}$$

where $\hat{\mathbf{z}}$ is the unit vector in the z-dimension. The diffraction efficiency, i.e., the fraction of incident light diffracted off the hologram, is given by the expression

$$\left|\frac{\mathcal{E}_d}{\mathcal{E}_{R'}}\right|^2 = \eta_0 \,\mathrm{sinc}^2\left(\frac{(\delta\boldsymbol{k}_d \cdot \hat{\mathbf{z}})\, L}{2\pi}\right), \tag{3.11}$$

where η_0 is a constant expressing the hologram strength. The quantity

$$\delta\boldsymbol{k}_d = \boldsymbol{k}_{R'} + \boldsymbol{K}_g - \boldsymbol{k}_d \tag{3.12}$$

is referred to as "Bragg-mismatch." If $\delta\boldsymbol{k}_d = 0$, the hologram is said to be "Bragg-matched," and the diffraction efficiency is maximum, equal to η_0. An equivalent expression for the Bragg-matching condition is

$$|\,\boldsymbol{K}_g + \boldsymbol{k}_{R'}\,| = k_{R'}. \tag{3.13}$$

The geometrical interpretation of this condition is shown in Fig. 3.2.

If $\boldsymbol{k}_{R'}$ is changed so that $\delta\boldsymbol{k}_d \neq 0$, then the diffraction efficiency drops according to the sinc-type law of (3.11). The amount $\Delta\boldsymbol{k}_{R'}$ required for the sinc function to reach its first null is called "Bragg selectivity," because a new hologram may be superimposed using $\boldsymbol{k}_{R'}$ as reference. For the purposes of the current discussion, it

suffices to say that the *crosstalk* between the two holograms is minimal when $\Delta \boldsymbol{k}_{R'}$ becomes equal to the Bragg selectivity or an integer multiple thereof. We will provide later (section 4.1, see also [71, 72, 73]) a detailed calculation of crosstalk between information-bearing holograms.

The explicit expression for the diffracted field in the simple example of Fig. 3.1 is:

$$\mathcal{E}_d(\boldsymbol{r};\boldsymbol{k}_{R'}) = a_0\, L \times \frac{\operatorname{sinc}\left[\frac{L}{2\pi}\left(K_{gz} + \sqrt{k_{R'}^2 - k_{R'x}^2 - k_{R'y}^2} - \sqrt{k_{R'}^2 - (k_{R'x}+K_{gx})^2 - k_{R'y}^2}\right)\right]}{\sqrt{k_{R'}^2 - (k_{R'x}+K_{gx})^2 - k_{R'y}^2}} \times$$

$$\exp\left\{i\left[(k_{R'x}+K_{gx})\; x + k_{R'y}\, y + \sqrt{k_{R'}^2 - (k_{R'x}+K_{gx})^2 - k_{R'y}^2}\; z\right]\right\}, \qquad (3.14)$$

where $a_0 = \|\boldsymbol{a}(\boldsymbol{K}_g)\|$ and we used the shorthand notation $v_{ab} \equiv \boldsymbol{v}_a \cdot \hat{\mathbf{b}}$ for the components of vector $\boldsymbol{v}_a$. The grating vector is

$$\boldsymbol{K}_g = K_{gx}\hat{\mathbf{x}} + K_{gz}\hat{\mathbf{z}} \qquad (3.15)$$

(the y component is ommited without loss of generality because x and y are equivalent). From this condition we may obtain the Bragg selectivities for two important cases: (a) if $\Delta \boldsymbol{k}_{R'}$ is due to a change in angle of the reference beam (Fig. 3.2c), then the "angular Bragg selectivity" is:

$$\Delta\theta_{R'} = m\frac{\lambda\cos\theta_S}{L\sin(\theta_R+\theta_S)}, \qquad m = 1, 2, \ldots ; \qquad (3.16)$$

(b) if $\Delta \boldsymbol{k}_{R'}$ is due to a change in wavelength of the reference beam (Fig. 3.2d), then the "wavelength Bragg selectivity" is:

$$\Delta\lambda_{R'} = m\frac{\lambda^2\cos\theta_S}{2L\sin^2\frac{1}{2}(\theta_R+\theta_S)}, \qquad m = 1, 2, \ldots \qquad (3.17)$$

In Appendix II we calculate in detail the effect of crystal rotation in diffraction efficiency in the transmission geometry.

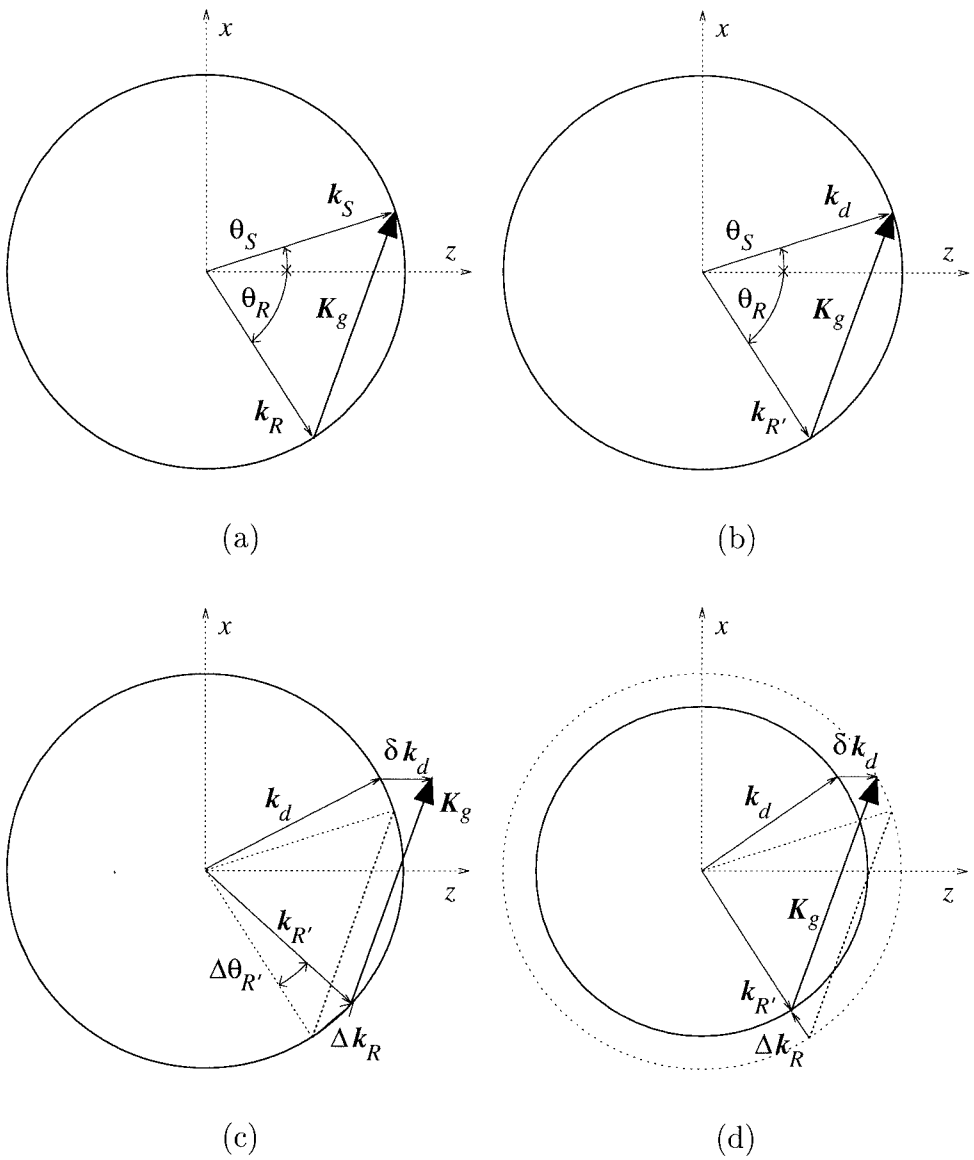


Figure 3.2: Illustration of Bragg-diffraction on the k-sphere: (a) recording of grating $\boldsymbol{K}_g$ by plane waves with wave-vectors $\boldsymbol{k}_R$, $\boldsymbol{k}_S$; (b) reconstruction by Bragg-matched beam with $\boldsymbol{k}_{R'} = \boldsymbol{k}_R$; (c) reconstruction by beam rotated by $\Delta\theta_{R'}$ (angle-multiplexing); (d) reconstruction by beam at wavelength detuned by $\Delta\lambda_{R'}$ (wavelength-multiplexing).

In volume holography, the Bragg selectivity is used to record multiple overlapping holograms sharing the same material volume, in other words for "hologram multiplexing." According to what has been described so far, this is achieved by changing the angle [36] or the wavelength [37, 38] of the reference beam by an amount equal to the respective selectivity. For typical parameters $\theta_R = \theta_S = 30°$, $L = 5$ mm, $\lambda = 488$ nm we obtain $\Delta\theta_R = 5.6° \times 10^{-3}$, and $\Delta\lambda_R = 8.2 \times 10^{-2}$ nm. For example, if the angular range yielded by typical lenses is 20°, or if the range of a tunable laser source is 300 nm, we obtain $M_\theta \approx 3,500$, and $M_\lambda \approx 3,600$ for the number of holograms that can be stored, respectively. Assuming that each hologram contains approximately 1 Mbit of information, and that the area occupied by the holograms is 4 mm^2 (2μm pixel size), we obtain the volume density of the holographic memory as $\approx$18 Gbits/cm^3.

Phase-code multiplexing [39, 74] is directly related to angle multiplexing in the sense that instead of using one plane wave reference at a time, one uses all of them at once, observing the Bragg-limited angular separation. The phases of the reference components implement some set of orthogonal functions, e.g., Walsh-Hadamard codes. Upon reconstruction, each member of the orthogonal reference set reconstructs its own hologram; orthogonality serves to eliminate all other reconstructions, yielding minimal crosstalk [75]. The maximum number of superimposed holograms using this method is equal to the order M_ϕ of the system of orthogonal functions. Bragg mismatch is employed to eliminate multiple reconstructions due to the multiple reference components. Therefore M_ϕ is limited by the number of beams angularly separated by $\Delta\theta_R$ (eq. 3.16) that fit in the aperture of the reference imaging system. In that sense, phase code multiplexing offers the same capacity as angle multiplexing ($M_\phi = M_\theta$) for the same optics. In practice, the space–bandwidth product of the available SLM's currently limits M_ϕ to $\leq 2,000$.

The capacity offered by holographic storage can be further augmented by combining angle and wavelength multiplexing [76], or either one of the two techniques with methods that do not utilize Bragg mismatch [40]. These methods are based on the property of reconstructed holograms to follow the motion of the reference beam when

Bragg mismatch is not present. For example, rotating the reference beam around the optical axis causes the reconstruction to rotate similarly until either it becomes Bragg–mismatched or moves out of the detector plane (which one of the two occurs first is determined by the spatial signal bandwidth). This effect is utilized in the method of peristrophic multiplexing [41] (in Greek, "peristrophic" means rotational). A more detailed discussion on the Bragg selectivity properties of peristrophic multiplexing is given in Appendix II of this section. Recently, surface storage density of 10bits/μm^2 without any observed errors in the reconstructions was demonstrated using combination of angle and peristrophic multiplexing [77]. The thickness of the material used for this experiment was 100μm, and volume density of 100 Gbits/cm^3 was achieved.

Appendix I: Validity of the paraxial approximation in volume diffraction

The electric field $\boldsymbol{E}$ satisfies the wave equation

$$\nabla\left(\nabla\cdot\boldsymbol{E}\right)-\nabla^2\boldsymbol{E}-\omega^2\,\mu_0\,\epsilon\,\boldsymbol{E}\;=\;0, \tag{3.18}$$

$$\epsilon(\boldsymbol{r}')=\epsilon_0+\Delta\epsilon_+(\boldsymbol{r}'). \tag{3.19}$$

The perturbation of the dielectric index inside the volume holographic material is expressed in Fourier space as

$$\Delta\epsilon_+\left(\boldsymbol{r}\right)=\Delta\epsilon_1\left(\boldsymbol{r}\right)e^{\,i\boldsymbol{K}_{g0}\cdot\boldsymbol{r}}=\int_{\mathcal{V}_{\mathbf{K}}}\Delta\tilde{\epsilon}\left(\boldsymbol{K}_g\right)e^{i\boldsymbol{K}_g\cdot\boldsymbol{r}}\mathrm{d}^3\boldsymbol{K}_g. \tag{3.20}$$

where $\boldsymbol{K}_{g0}$ is the carrier wave-vector and $\mathcal{V}_{\mathbf{K}}$ is the portion of Fourier space where the hologram spectrum has significant values. The bandlimited nature of the hologram is expressed by the conditions

$$\boldsymbol{K}_g\;-\;\boldsymbol{K}_{g0}\;\equiv\;\Delta\boldsymbol{K}_g,\qquad|\Delta\boldsymbol{K}_g|<<|\boldsymbol{K}_{g0}|\qquad\forall\;\Delta\boldsymbol{K}_g\in\mathcal{V}_{\mathbf{K}}.$$

The $\boldsymbol{A}(\boldsymbol{k}_{R'}, \boldsymbol{k}_d)$ factor is computed easily as superposition of the individual Fourier components of the perturbation, and is given by

$$\boldsymbol{A}(\boldsymbol{k}_{R'}, \boldsymbol{k}_d) = \int_{\mathcal{V}_{\mathbf{K}}} \boldsymbol{a}(\boldsymbol{K}_g)\, \Delta\tilde{\epsilon}(\boldsymbol{K}_g) \left(\int_{\mathcal{V}} \mathrm{e}^{i(\boldsymbol{k}_{R'} + \boldsymbol{K}_g - \boldsymbol{k}_d)\cdot \boldsymbol{r'}} \mathrm{d}^3\boldsymbol{r'} \right) \mathrm{d}^3\boldsymbol{K}_g, \tag{3.21}$$

where

$$\begin{aligned} \boldsymbol{a}(\boldsymbol{K}_g) &= \frac{1}{2i\epsilon_0 k_{dz}} \left[(\boldsymbol{K}_{g0} + \Delta\boldsymbol{K}_g) \cdot \hat{\mathbf{e}}_\mathbf{i} \right] (\boldsymbol{k}_{R'} + \boldsymbol{K}_{g0} + \Delta\boldsymbol{K}_g) - \omega^2 \mu_0 \epsilon_0 \hat{\mathbf{e}}_\mathbf{i}, \\ &= \boldsymbol{a}_0 + \frac{\boldsymbol{K}_{g0} \cdot \hat{\mathbf{e}}_\mathbf{i}}{\epsilon_0} \Delta\boldsymbol{K}_g + \frac{\Delta\boldsymbol{K}_g \cdot \hat{\mathbf{e}}_\mathbf{i}}{\epsilon_0} (\boldsymbol{k}_{R'} + \boldsymbol{K}_{g0}) \\ &\quad + \mathcal{O}\left(|\Delta\boldsymbol{K}_g|^2 \right) \qquad (3.22) \\ \boldsymbol{a}_0 &= \frac{1}{2i\epsilon_0 k_{dz}} (\boldsymbol{K}_{g0} \cdot \hat{\mathbf{e}}_\mathbf{i}) (\boldsymbol{k}_{R'} + \boldsymbol{K}_{g0}) - \omega^2 \mu_0 \epsilon_0 \hat{\mathbf{e}}_\mathbf{i}. \qquad (3.23) \end{aligned}$$

Each term in (3.22) can be integrated separately, yielding

$$\begin{aligned} \mathbf{A}(\boldsymbol{k}_{R'}, \boldsymbol{k}_d) &= \boldsymbol{a}_0 \int_{\mathcal{V}} \Delta\epsilon(\boldsymbol{r'})\, \mathrm{e}^{i(\boldsymbol{k}_i - \boldsymbol{k}_d)\cdot \boldsymbol{r'}} d^3\boldsymbol{r'} - \\ &\quad \frac{\boldsymbol{K}_{g0} \cdot \hat{\mathbf{e}}_\mathbf{i}}{2\epsilon_0 k_{dz}} \int_{\mathcal{V}} \left(\mathrm{e}^{i\boldsymbol{K}_{g0}\cdot \boldsymbol{r'}} \nabla\Delta\epsilon_0(\boldsymbol{r'}) \right) \mathrm{e}^{i(\boldsymbol{k}_{R'} - \boldsymbol{k}_d)\cdot \boldsymbol{r'}} \mathrm{d}^3\boldsymbol{r'} - \\ &\quad \frac{(\boldsymbol{k}_{R'} + \boldsymbol{K}_{g0})}{2\epsilon_0 k_{dz}} \int_{\mathcal{V}} \left(\mathrm{e}^{i\boldsymbol{K}_{g0}\cdot \boldsymbol{r'}} \nabla\Delta\epsilon_0(\boldsymbol{r'}) \cdot \hat{\mathbf{e}}_\mathbf{i} \right) \mathrm{e}^{i(\boldsymbol{k}_{R'} - \boldsymbol{k}_d)\cdot \boldsymbol{r'}} \mathrm{d}^3\boldsymbol{r'} \\ &\quad + \mathcal{O}\left(|\Delta\boldsymbol{K}_g|^2 \right). \qquad (3.24) \end{aligned}$$

The first term in (3.24) is the approximate solution of (3.5), while the remaining terms express the first-order corrections, which depend on the *gradient* of the perturbation. If we further assume that the gradient is bounded above, as in

$$| \nabla\Delta\epsilon_0(\boldsymbol{r'}) | \leq | \boldsymbol{\Delta}_0 | \qquad \forall\, \boldsymbol{r'} \in \mathcal{V},$$

then we obtain an upper bound on the first-order error in (3.5):

$$|\,\delta \boldsymbol{A}\,(\boldsymbol{k}_{R'}, \boldsymbol{k}_d)\,| \;\leq \\ \frac{|\Delta \boldsymbol{K}_g|_{\max}}{2\,\epsilon_0\,k_{dz}}\;\{\;|\,\boldsymbol{K}_{g0}\cdot\hat{\mathbf{e}}_{\mathbf{i}}\,|\,|\,\boldsymbol{\Delta}_0\,|+|\,\boldsymbol{k}_{R'}+\boldsymbol{K}_{g0}\,|\;[\,\max|\nabla\Delta\epsilon_0\cdot\hat{\mathbf{e}}_{\mathbf{i}}|\,]\}\,. \quad (3.25)$$

Note that the first-order distortion is eliminated if the conditions $\boldsymbol{K}_{g0} \perp \hat{\mathbf{e}}_{\mathbf{i}}$ and $\nabla\Delta\epsilon_0 \perp \hat{\mathbf{e}}_{\mathbf{i}}$ are satisfied.

Appendix II: Bragg mismatch effects from crystal rotation

As a more complete example of calculating diffraction from Bragg-mismatched volume gratings, we now consider the effect of rotation on volume diffraction. Consider a volume hologram of thickness L in the z direction, infinite in the x, y directions, expressed as

$$\Delta\epsilon\,(\boldsymbol{r}') = \Delta\epsilon_0 \mathrm{rect}\left(\frac{z'}{L}\right) \mathrm{e}^{i\boldsymbol{K}_g\cdot\boldsymbol{r}'}. \quad (3.26)$$

The grating vector is given by (3.15) and the diffracted field by (3.14) according to the theory of the main part of this section. We will now consider the three possible rotations in sequence.

(a) z-rotation *(Peristrophic multiplexing)* [41]

To treat rotation around the z-axis, we express the readout reference wavevector as

$$\boldsymbol{k}_{R'} = k_{R'z}\hat{\mathbf{z}} + \boldsymbol{k}_{\perp}.$$

Assume that $\boldsymbol{k}_{R'}$ is rotated by a small angle ψ_z relative to the Bragg-matched position. This small rotation causes no change in $k_{R'z}$. The transverse components change by

$$\Delta k_{R'x} \approx -k_{R'y}\,\psi_z \qquad \Delta k_{R'y} \approx k_{R'x}\,\psi_z.$$

Substituting in (3.14), we find that the diffracted intensity varies with ψ_z, to order

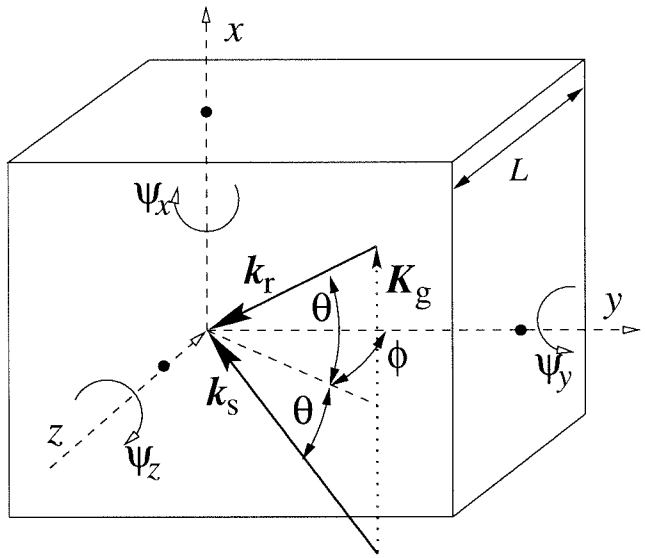


Figure 3.3: Symmetric recording geometry.

$\mathcal{O}(\psi_z)$, like

$$I_{\text{diffr}} \propto \text{sinc}^2\left(\frac{L}{2\pi}\,\frac{k_{R'y}\,K_{gx}}{k_{R'z}+K_{gz}}\,\psi_z\right). \tag{3.27}$$

The first Bragg null occurs when the argument to the sinc in the above expression becomes one; therefore, the selectivity for rotation around z is

$$\Delta\psi_z = \frac{2\pi}{L}\left|\frac{k_{R'z}+K_{gz}}{k_{R'y}\,K_{gx}}\right|. \tag{3.28}$$

This calculation is valid only if the denominator in (3.28) is non-zero. In the singular case $k_{R'y}=0$ (i.e., when the input wave vector and the grating vector are co-planar), then keeping $\mathcal{O}\,({\psi_z}^2)$ terms in the Taylor expansion yields

$$\Delta\psi_z = \left\{\frac{2\pi}{L}\left|\frac{k_{R'z}+K_{gz}}{k_{R'x}\,K_{gx}}\right|\right\}^{1/2}. \tag{3.29}$$

It is interesting to examine the transition between the usual Bragg mismatch effect and the singular case of $\boldsymbol{k}_{R'}$, $\boldsymbol{K}_g$ co-planar. For the moment we will consider the symmetric geometry of Figure 3.3 and we will generalize later. The recording

reference, signal and grating wave-vectors are

$$\boldsymbol{k}_R = k\left(-\sin\theta, \cos\theta\sin\phi, \cos\theta\cos\phi\right) \tag{3.30}$$

$$\boldsymbol{k}_S = k\left(\sin\theta, \cos\theta\sin\phi, \cos\theta\cos\phi\right) \tag{3.31}$$

$$\boldsymbol{K}_g = \boldsymbol{k}_S - \boldsymbol{k}_R = k\left(2\sin\theta, 0, 0\right). \tag{3.32}$$

The peristrophic Bragg-mismatch for the symmetric geometry, keeping two orders of ψ_z, is

$$\Delta k_z = \frac{k_{R'y}K_{gx}}{k_{dz}^2}\psi_z + \frac{k_{R'x}K_{gx}}{2k_{dz}^2}\left(1 - \frac{k_{R'y}^2 K_{gx}}{k_{dz}^2 k_{R'x}}\right)\psi_z^2. \tag{3.33}$$

By requiring $L\Delta k_z/(2\pi) = \pm 1$ we obtain for the Bragg mismatch the following quadratic condition:

$$A\left(\Delta\psi_z\right)^2 - B\Delta\psi_z \pm C = 0, \qquad \text{where} \tag{3.34}$$

$$A = \frac{\tan^2\theta}{\cos^2\phi}\left(1 + 2\tan^2\phi\right), \tag{3.35}$$

$$B = \frac{2\tan\theta\tan\phi}{\cos\phi}, \tag{3.36}$$

$$C = \frac{\lambda}{L\cos\theta\cos\phi}. \tag{3.37}$$

If ϕ is not negligible, the approximate solution to the quadratic equation (3.34) is

$$|\Delta\psi_z| \approx \frac{\lambda\cot\phi}{2L\sin\theta}, \tag{3.38}$$

in agreement with (3.28), as can be verified easily. On the other hand, if $\phi \approx 0$, the linear term in (3.34) vanishes and we obtain

$$|\Delta\psi_z| \approx \sqrt{\frac{\lambda\cos\theta}{L\sin^2\theta}}, \tag{3.39}$$

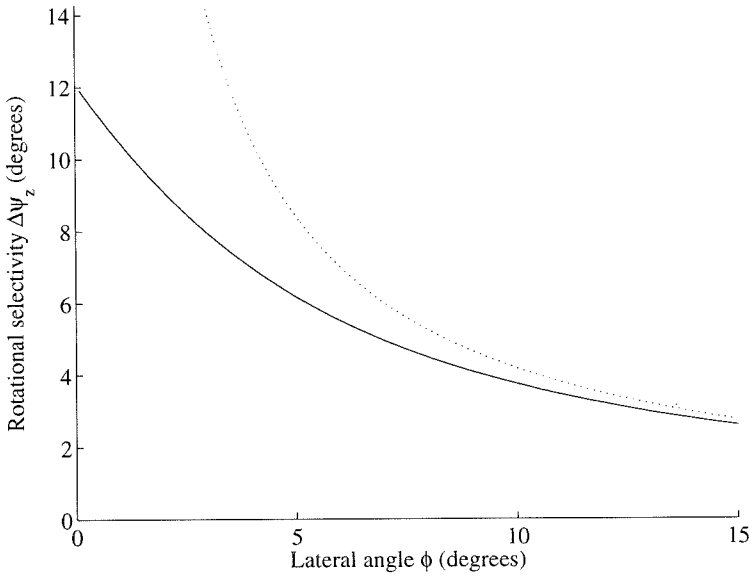


Figure 3.4: Transition from non-singular to singular Bragg selectivity (solid curve) as $\phi \to 0$ for $\lambda = 488$ nm, $L = 38\mu$m, $\theta = 30°$. The dotted curve is the Bragg selectivity approximation (3.28), which breaks down for small ϕ.

in agreement with (3.29). The transition is illustrated in Figure 3.4 using both the exact solution of (3.34) and the approximation (3.28).

Now consider the case of a reflection grating $\boldsymbol{K}_g = K\hat{\mathbf{z}}$. We can see that then all terms in (3.33) vanish, which means that relative rotation of the hologram around z does not cause Bragg mismatch. The physical interpretation is that the hologram is symmetric around the axis of rotation, and therefore does not "sense" the rotating reference. This is a case of a "degenerate" grating (for the particular type of rotation).

(b) x-rotation

Similarly to the case of z-rotation, and omitting the details, we obtain

$$\Delta\psi_x = \frac{2\pi}{L}\left|\frac{k_{R'z} + K_{gz}}{k_{R'y}\ K_{gz}}\right|. \tag{3.40}$$

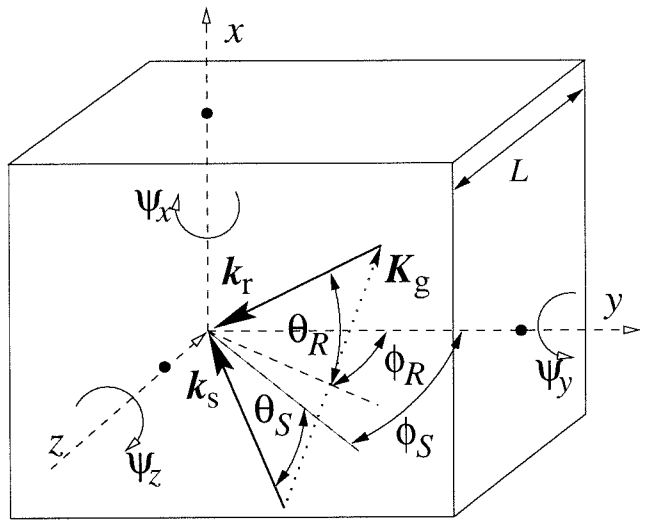


Figure 3.5: General (asymmetric) recording geometry for a grating lying on the xz-plane.

The selectivity becomes singular when $k_{R'y} = 0$, and is given by

$$\Delta\psi_x = \left\{ \frac{2\pi}{L} \left| \frac{k_{R'z} + K_{gz}}{k_{R'z}\ K_{gz}} \right| \right\}^{1/2}. \tag{3.41}$$

The grating becomes degenerate when $K_{gz} = 0$.

(c) y-rotation

In this case we obtain

$$\Delta\psi_y = \frac{2\pi}{L} \left| \frac{k_{R'z} + K_{gz}}{k_{R'z}\ K_{gx} - k_{R'x}\ K_{gz}} \right|. \tag{3.42}$$

The singularity occurs when $k_{R'x} \approx K_{gx} \approx 0$, and the selectivity is then

$$\Delta\psi_y = \sqrt{\frac{\lambda}{L}}. \tag{3.43}$$

Rotational selectivities for the asymmetric geometry

We conclude this Appendix with the expressions for the rotational Bragg selectivities of the general asymmetric geometry of Figure 3.5. The derivations are straight-

forward from the preceding theory.

z-Rotation

$$\Delta\psi_z = \frac{\lambda}{L}\,\frac{\cos\theta_R\cot\phi_R + \cos\theta_S\sin(\phi_S-\phi_R)}{\cos\theta_R(\sin\theta_S+\sin\theta_R)} \tag{3.44}$$

Singularity ($\phi_R \approx \phi_S \approx 0$):

$$\Delta\psi_z = \left\{\frac{2\lambda}{L}\,\frac{\cos\theta_S}{\sin\theta_R(\sin\theta_S+\sin\theta_R)}\right\}^{1/2} \tag{3.45}$$

Degeneracy condition: $K_{gx} = 0$.

x-Rotation

$$\Delta\psi_x = \frac{\lambda}{L}\,\frac{\cos\theta_R\cos\phi_R + \cos\theta_S\sin\phi_R\sin(\phi_S-\phi_R)}{\cos^2\theta_S\sin^2\phi_R\sin(\phi_S-\phi_R)} \tag{3.46}$$

Singularity ($\phi_R \approx \phi_S \approx 0$):

$$\Delta\psi_x = \left\{\frac{\lambda}{L}\,\frac{\cos\theta_S}{\cos\theta_R(\cos\theta_R-\cos\theta_S)}\right\}^{1/2} \tag{3.47}$$

Degeneracy condition: $K_{gz} = 0$.

y-Rotation

$$\Delta\psi_y = \frac{\lambda}{L}\,\frac{\cos\theta_R\cos\phi_R + \cos\theta_S\sin\phi_R\sin(\phi_S-\phi_R)}{\cos\theta_R\cos\phi_R(\sin\theta_S+\sin\theta_R) + \sin\theta_R\cos\theta_S\sin\phi_R\sin(\phi_S-\phi_R)} \tag{3.48}$$

Singularity:

$$\Delta\psi_y = \left\{\frac{\lambda}{L}\right\}^{1/2} \tag{3.49}$$

No degeneracy.

3.2 Array multiplexing

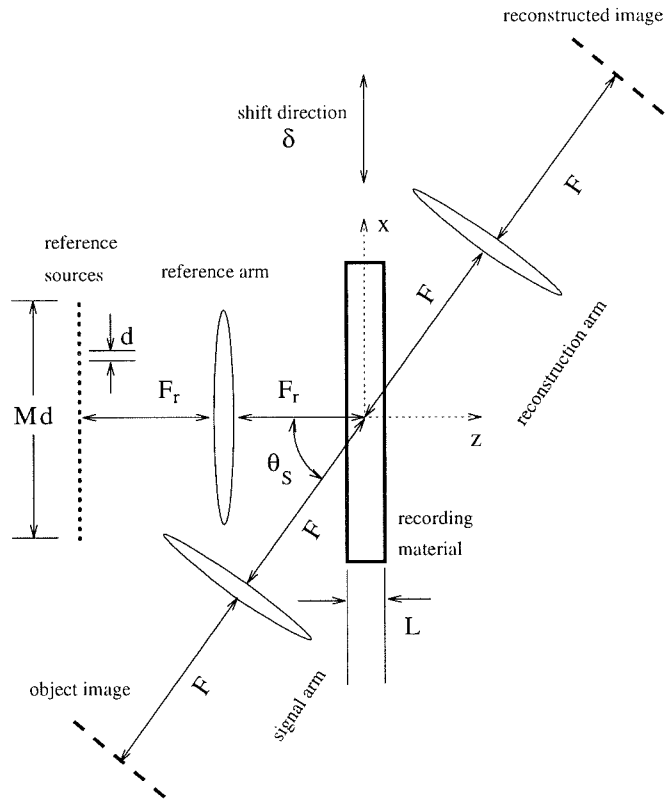


Figure 3.6: Geometry for shift multiplexing in the Fourier Plane.

In this section, we introduce array multiplexing [42] as a method for holographic storage. To implement this method, the reference beam must consist of a spectrum of plane waves (similar to phase code multiplexing [39, 78], for example). Multiplexing is achieved by shifting the recording medium with respect to the signal and reference beams. Alternatively the two beams can be translated in tandem with respect to the stationary medium.

The geometry for array multiplexing is shown in Fig. 3.6 for the case of storing Fourier transform holograms. The reference originates from an array of M point sources located in the front focal plane of a Fourier lens, and centered around the optical axis z. The lens transforms the field into a fan of M plane waves. The angular separation is uniform, given by $\Delta\theta \approx d/F_r$ where d is the distance between successive

point sources and F_r is the focal length. Thus the angle of incidence of the m–th component is

$$\theta_m \approx \left(m - \frac{M-1}{2}\right)\Delta\theta \qquad m = 0, \ldots, M-1 \tag{3.50}$$

The angle of incidence of the central component of the signal with respect to the z–axis is denoted by θ_S.

Because the reference consists of M plane waves, we can think of the recording as consisting of M separate holograms recorded simultaneously. Upon reconstruction, each plane wave in the reference fan reads out not only the hologram it recorded, but also all the holograms recorded by the other plane waves of the reference fan. These additional reconstructions, or "ghosts," produce images that are shifted with respect to the primary reconstruction, due to the change in read–out angle relative to the recording angle. The ghosts are Bragg mismatched by an amount roughly proportional to the angular separation between the plane wave component that originally recorded the hologram and the component that is reconstructing it. For the hologram recorded between the central signal component and the $m = 0$-th reference component, the amount of Bragg mismatch is $\Delta k_z = 2\pi l \tan\theta_S \Delta\theta/\lambda$ when read out by the $\pm l$-th reference component. The same relation holds approximately for the other holograms. The diffraction efficiency of these Bragg mismatched holograms is proportional to

$$\eta(\Delta k) = \mathrm{sinc}^2\left(\frac{\Delta k_z\, L}{2\pi}\right) \tag{3.51}$$

where $\mathrm{sinc}\,(x) = \sin\pi x/(\pi x)$ and L is the thickness of the recording medium. It follows that by choosing the angular separation $\Delta\theta$ between the reference components such that the sinc function of (3.51) vanishes, the ghosts will be eliminated, leaving a clean reconstruction. From (3.51) the required separation is:

$$\Delta\theta \approx \frac{\lambda}{L\tan\theta_S} \tag{3.52}$$

Having eliminated the ghosts, we now examine what happens to the diffracted light if the hologram is shifted by a distance δ in the x–direction (see Fig. 3.6). The diffracted field $\mathcal{E}_d$ is obtained by multiplying the illuminating reference (consisting of M plane waves) by the expression for the M recorded holograms shifted by δ. For a single plane wave signal beam of incidence angle θ_S, we have:

$$\begin{aligned}
\mathcal{E}_d &= \sum_m \exp\left\{ i2\pi \frac{m\Delta\theta x}{\lambda} \right\} \times \\
&\quad \sum_{m'} \exp\left\{ -i2\pi \frac{m'\Delta\theta(x-\delta)}{\lambda} \right\} \exp\left\{ i2\pi \sin\theta_S \frac{x-\delta}{\lambda} \right\} \qquad (3.53) \\
&\approx \left(\sum_m \exp\left\{ i2\pi \frac{m\Delta\theta\delta}{\lambda} \right\} \right) \exp\left\{ i2\pi \sin\theta_S \frac{x-\delta}{\lambda} \right\} \qquad (3.54)
\end{aligned}$$

The three-dimensional nature of the hologram (i.e., the z dependence) serves to eliminate the cross–terms $m \neq m'$ (ghosts) from the double summation. When a signal with finite bandwidth is reconstructed, a detailed calculation (not given here) shows that the cross-terms are not eliminated completely, but the signal is still reconstructed with a high signal to noise ratio. From (3.54), the diffracted field consists of the reconstruction of the signal at angle θ_S, weighted by a sum leading to the familiar Helmholtz function [79], encountered often in the theory of antenna arrays. For notational simplicity, we define:

$$\mathrm{ar}(\omega; m) = \frac{\sin(m\pi\omega)}{l\sin(\pi\omega)}. \qquad (3.55)$$

Then the diffracted intensity as function of shift is:

$$I(\delta) \propto \mathrm{ar}^2\left(\frac{\delta\Delta\theta}{\lambda} \; ; \; M \right) \qquad (3.56)$$

The zeros of the Helmholtz function occur at

$$\delta_l = l\frac{\lambda}{M\Delta\theta}, \qquad l = 1, \ldots, M-1. \qquad (3.57)$$

Multiplexing is performed by recording each hologram with a shift $\delta_1 = \lambda/M\Delta\theta$ with

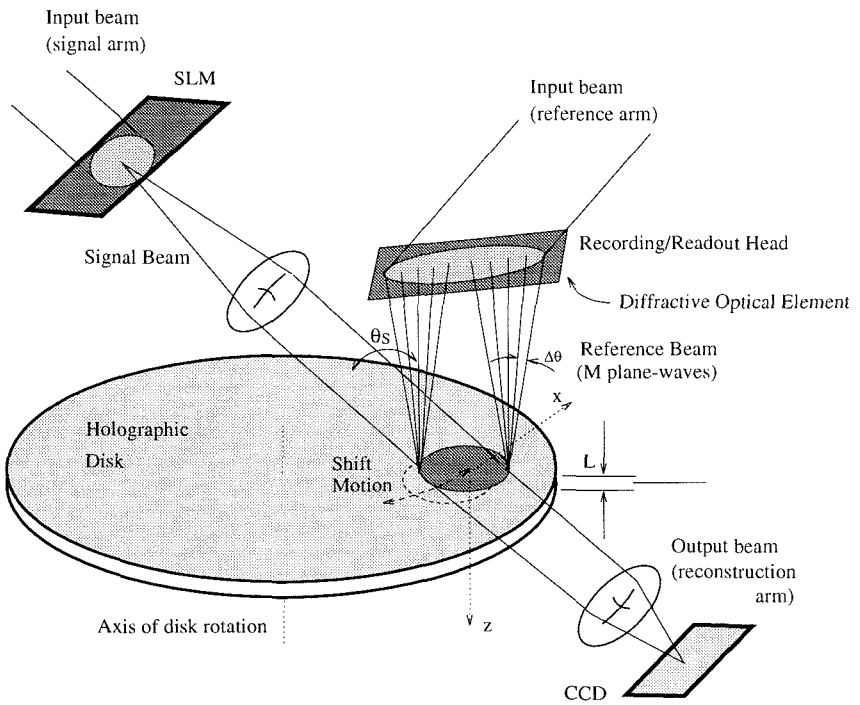


Figure 3.7: Holographic 3–D disk with array-multiplexed holograms.

respect to its two neighbors. Because of the periodicity of the array function, at maximum M holograms can be superimposed on the same location. The period is:

$$\delta_M = \frac{\lambda}{\Delta\theta}. \tag{3.58}$$

The array multiplexing method is particularly well suited for the implementation of holographic 3-D disks [32, 44]. A schematic diagram is given in Fig. 3.7. The fan of reference waves, arranged along the tangential component of disk motion, is produced by a Diffractive Optical Element (DOE), optimized for the desired separation angle and equal intensities for the diffracted orders. Recording and accessing array-multiplexed holograms are readily implemented by simply using the disk rotation (which is already part of the system intended to allow accessing of information on different locations on the disk surface) in order to shift the disk until the correct location within a track is reached. Radial head motion (also present in optical disk

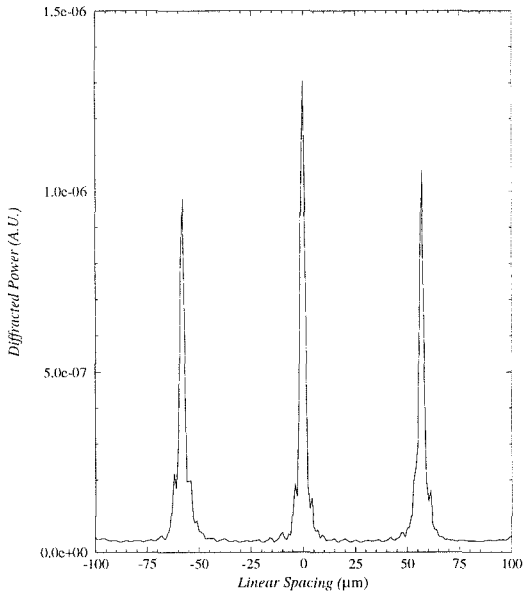


Figure 3.8: Experimental demonstration of the array function with a single hologram of a random bit pattern.

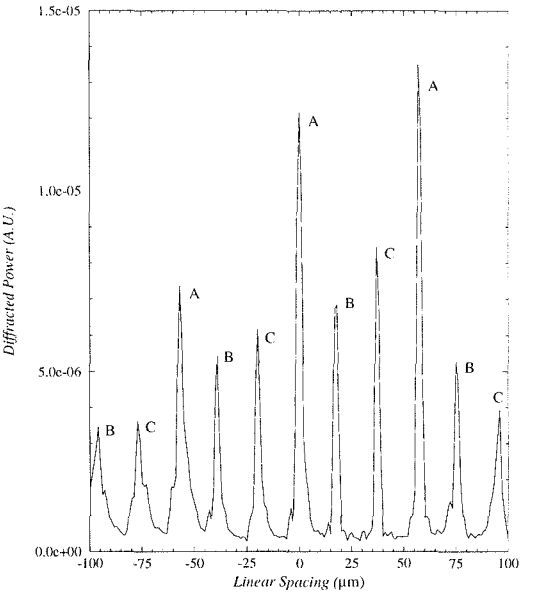


Figure 3.9: Multiplexing of three holograms (A, B, C) of random bit patterns using the shift method.

mechanisms) is used to access different tracks. This simplifies the design of the head since no additional components are required for selective readout. Comparing the array-multiplexed disk and the angle-multiplexed disk of Fig. 1.3, we see that the beam-steering mecahnism of the latter is replaced by a simple DOE in the former, making the system much more simple and robust. The design of the DOE, however, may be expensive.

The storage density $\mathcal{D}$ per unit area that we can achieve using the device of Fig. 3.7 is limited by the thickness dependent angular selectivity (eq. 3.52), the number of beams M allowed by the optics, the page size $N_p b$ (b is the pixel size and N_p the number of pixels) and the periodicity of the array function. An approximate formula for the density is:

$$\mathcal{D} = \frac{M \cos\theta_S}{b^2(1 + M\delta_1 \cos\theta_S / N_p b)}. \tag{3.59}$$

For $L = 100\mu$m and signal incidence angle $\theta_S = 30^o$, usage of F/1 optics allows $M = 100$ holograms. Then, for typical page parameters $N_p = 1000$, $b = 2\mu$m, eq. 3.59 yields $\mathcal{D} = 21.1\text{bits}/\mu\text{m}^2$.

Array multiplexing was demonstrated[2] using a reference fan of 20 plane waves angularly separated by 0.5°. The recording material was DuPont HRF-150 polymer of thickness $L = 38\mu$m. The DOE in this case was 20 plane wave holograms recorded on another sheet of polymer with a common reference, such that upon reconstruction they produced the desired plane wave fan. Thus the architecture was similar to Fig. 3.7 with the sheet of polymer acting as DOE.

The effect of shift on the reconstruction of a single hologram is shown in Fig. 3.8. The signal image was a 100×100 random bit pattern. For the particular parameters the theoretical shift selectivity is 2.8μm and the period is 55μm, in good agreement with the experiment. The reason for the deviation from the theoretically predicted periodicity is the finite transverse size of the recording region. Three holograms array–multiplexed with the same setup are shown in Fig. 3.9. Each hologram is reconstructed almost periodically, following its own array factor. Because of the very small thickness of the recording medium in this experiment, we used angular separation smaller than that predicted by eq. 3.52. Therefore the ghosts had to be filtered out in the Fourier plane.

3.3 Shift multiplexing

Shift multiplexing [42, 43] is a holographic storage method particularly suitable for holographic 3–D disks [32, 44]. The design of a shift multiplexed disk is shown in Fig. 3.10. It is similar to the architecture of the array-multiplexed disk except the reference is a spherical wave produced by a lens of high numerical aperture. The data is stored on the disk as a hologram recorded by the interference of the signal and the spherical reference.

As in the case of the array reference, the non-planar phase-front of the reference beam allows one to multiplex and selectively retrieve holograms simply by translating the disk relative to the recording head, as shown in the figure. The shift selectivity, i.e., the translation required to resolve shift multiplexed holograms, is typically in

[2]The experiments in Figs. 3.8 and 3.9 were conducted by Allen Pu.

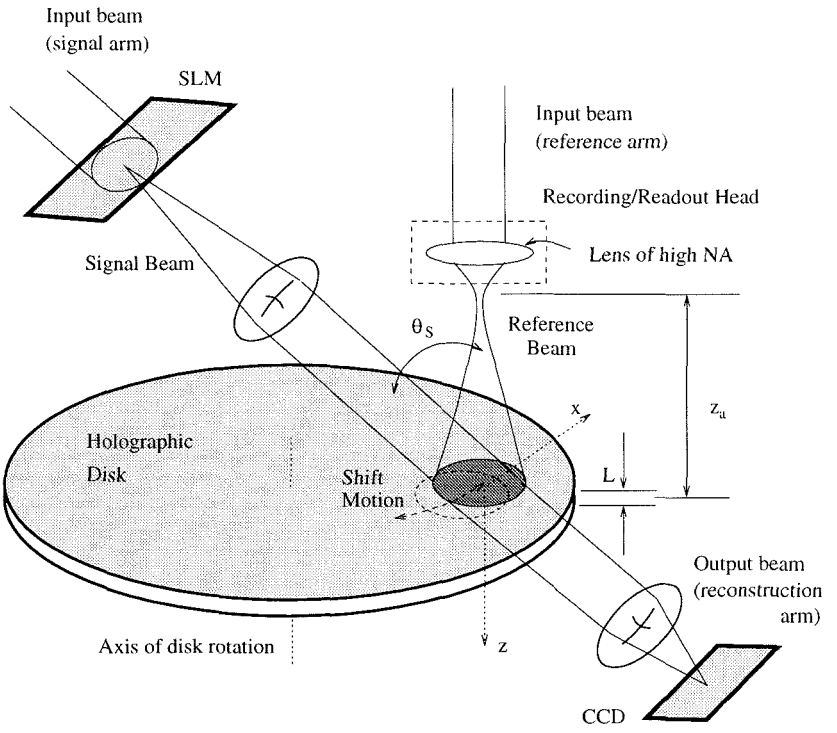


Figure 3.10: Holographic 3–D disk with shift-multiplexed holograms.

the order of a few microns, much less than the transverse size of the holograms (the latter is typically a few millimeters). In this way multiple overlapping holograms are superimposed. To selectively reconstruct holograms belonging to the same track, the disk is rotated relative to the stationary head. The head needs to move only in the radial direction to access different tracks on the disk. No additional multiplexing mechanism is needed. Since both disk rotation and radial head translation are an integral part of the optical disk configuration, a shift multiplexed disk is a very simple implementation.

In this section we derive the selectivity properties of shift-multiplexed holograms in the transmission, 90°, and fractal geometries (sections 3.3.1, 3.3.2, and sec:shift-fractal respectively). Further analysis of the properties of shift-multiplexed holographic 3–D disks is given in the next chapter.

3.3.1 Transmission geometry

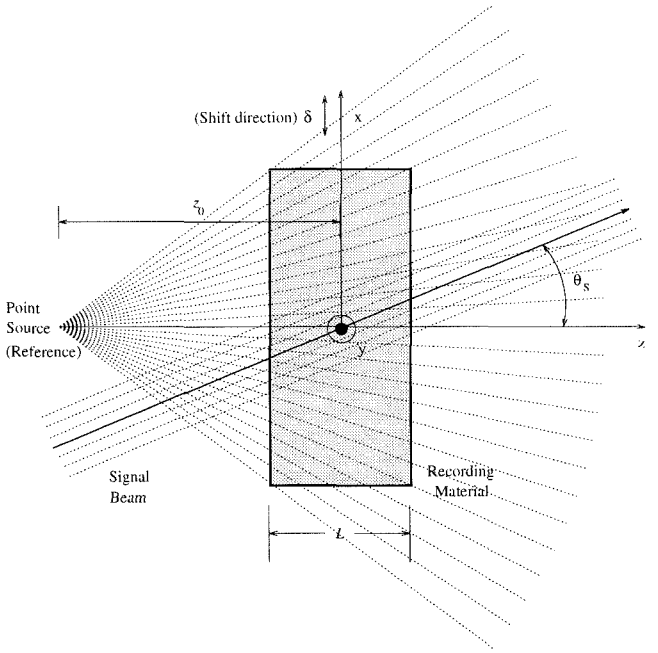


Figure 3.11: Geometry for shift multiplexing using a spherical reference wave.

The use of spherical reference beams in volume holography was treated in [80, 81, 82, 83]. In [80], a spherical reference was used for a holographic correlator, and the shift invariance curves were obtained theoretically and experimentally. Here we use a similar approach to derive the shift selectivity of shift multiplexed memories.

The geometry for shift multiplexing using spherical waves is shown in Fig. 3.11. The hologram is recorded in the region $|z| < L/2$ and is assumed infinite in the transverse directions x, y. The spherical reference wave is produced by a spherical lens of high numerical aperture. The focus is located at $z = -z_0$. The expression for the reference beam in the chosen system of coordinates, and under the paraxial approximation, is:

$$R(x,y,z) = \frac{1}{i\lambda(z+z_0)} \exp\left\{i2\pi\frac{z+z_0}{\lambda}\right\} \exp\left\{i\pi\frac{x^2+y^2}{\lambda(z+z_0)}\right\}. \tag{3.60}$$

We consider a plane wave component of the signal beam propagating on the xz-

plane, making angle θ_S with the z axis, expressed as:

$$S(x,z) = \exp\left\{i2\pi u_S \frac{x}{\lambda} + i2\pi\left(1 - \frac{u_S^2}{2}\right)\frac{z}{\lambda}\right\}, \tag{3.61}$$

where $u_S \equiv \sin\theta_S \approx \theta_S << 1$ (paraxial approximation). If we neglect the variation of the modulation depth throughout the hologram due to the defocusing of the spherical wave, then the hologram can be expressed by the term $R^*(x,y,z)S(x,z)$ in the resulting interference pattern. We now consider the expression for the field diffracted from a thin layer of the hologram located at z using a displaced reference beam $R(x-\delta,y,z)$:

$$\begin{aligned} R(x-\delta,y,z)R^*(x,y,z)S(x,z) = \\ \exp\left\{-i\pi\frac{2\delta x}{\lambda(z+z_0)}\right\}\exp\left\{i\pi\frac{\delta^2}{\lambda(z+z_0)}\right\}\times \\ \exp\left\{i2\pi u_S\frac{x}{\lambda}\right\}\exp\left\{i2\pi\left(1-\frac{u_S^2}{2}\right)\frac{z}{\lambda}\right\}. \end{aligned} \tag{3.62}$$

The signal beam is reconstructed if $\delta = 0$. For $\delta \neq 0$, the first term in (3.62) causes the reconstruction to deviate angularly with respect to the original signal $S(x,z)$ by an amount

$$\Delta\theta_S \approx \frac{\delta}{(z+z_0)\cos\theta_S}. \tag{3.63}$$

Since this angular deviation has a z dependence, reconstructions coming from successive thin slices of the hologram are phase mismatched. The amount of shift δ required to exactly cancel the reconstruction is calculated in the Appendix to this section (under the paraxial, Born, and constant modulation depth approximations) and it is given by:

$$\delta_{\text{Bragg}} = \frac{\lambda z_0}{L u_S}. \tag{3.64}$$

It is interesting that, in the geometry of Fig. 3.11, if the reference were a plane wave

incident along the z axis instead of the spherical wave, then the angular selectivity would be

$$\Delta\theta = \frac{\lambda}{L\tan\theta_S} \approx \frac{\lambda}{Lu_S}. \tag{3.65}$$

Thus we obtain the useful formula

$$\delta_{\text{Bragg}} = \Delta\theta \times z_0. \tag{3.66}$$

The finite spot size $\Delta x = \lambda/2(\text{NA})$ of a truncated spherical wave introduces ambiguity in the location of the point source with respect to the hologram. This ambiguity must be added to the shift selectivity, giving the final expression:

$$\begin{aligned} \delta &= \delta_{\text{Bragg}} + \Delta x \\ &= \frac{\lambda z_0}{L\tan\theta_S} + \frac{\lambda}{2(\text{NA})}. \end{aligned} \tag{3.67}$$

So far we assumed a holographic medium with index of refraction equal to 1. Unless an index–matching liquid is used, the change in refraction index n_0 at the interface of the holographic material causes the apparent location of the point source (as seen by an observer inside the holographic medium) to move away from the hologram. If we let z_{a} denote the distance of the point source from the center of the holographic material, measured in air, then the apparent z_0 relates to z_{a} (paraxially) as:

$$z_0 - \frac{L}{2} = n_0\left(z_{\text{a}} - \frac{L}{2}\right). \tag{3.68}$$

Therefore the modified selectivity equation is:

$$\delta = \frac{\lambda_0\left[z_{\text{a}} - \left(1 - \dfrac{1}{n_0}\right)\dfrac{L}{2}\right]}{L\tan\theta'_S} + \frac{\lambda_0}{2\,(\text{NA})} \tag{3.69}$$

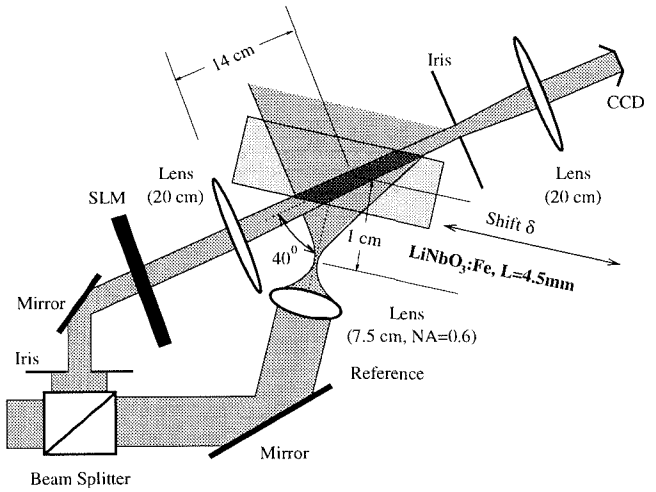


Figure 3.12: Experimental set–up for the demonstration of shift multiplexing (not drawn to scale).

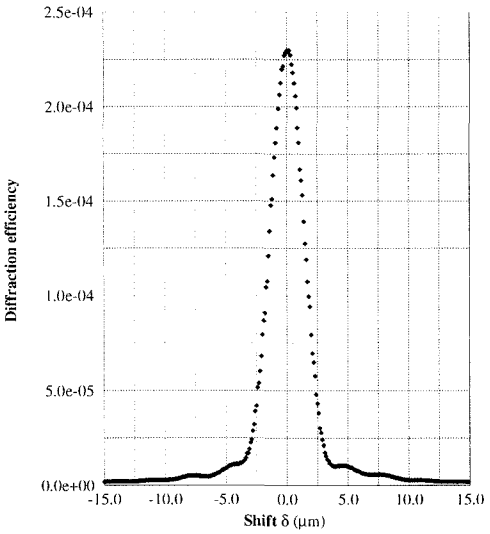


Figure 3.13: Experimental selectivity curve (diffraction efficiency η versus shift δ). The parameters of the experiment are given in Fig. 3.12.

where λ_0 denotes the wavelength of light in vacuum, and θ'_S is the angle of incidence of the signal inside the material, determined from Snell's law.

The experimental geometry used for all the shift multiplexing experiments described in this paper is shown in Fig. 3.12. The experimental parameters were $\lambda_0 = 488$ nm, $L = 4.5$ mm, $\theta_S = 40°$ (measured outside the crystal), $z_a = 1$ cm (distance from focus of the spherical reference to the center of the crystal, measured in air), and numerical aperture NA=0.6. The recording material (iron-doped $LiNbO_3$) has index of refraction $n_0 \approx 2.24$. The signal was a chessboard pattern, recorded as a Fresnel region hologram. The size of each square in the chessboard at the SLM plane was approximately 0.5 mm. For the parameters used in the experiment, eq. (3.69) yields $\delta = 3.58\mu$m. The experimental selectivity curve is shown in Fig. 3.13. The first null occurred at approximately $3.7 \pm 0.2\mu$m (the margin of error is mainly due to stage inaccuracy and backlash), deviating by 3.6% from the theoretical prediction.

Appendix: Derivation of the Bragg shift selectivity

In this appendix we derive the diffraction efficiency of a spherical volume hologram as a function of the shift δ of the reference relative to the hologram. We consider

again the geometry of Fig. 3.11. Under the Born approximation, the diffracted field at the observation point $\boldsymbol{r}_{\rm p}$ is given by the 3–D convolution integral (3.7) which is repeated here with some changes in notation:

$$\mathcal{E}_d(\boldsymbol{r}_{\rm p}) = \int_{\mathcal{V}} \mathcal{E}_i(\boldsymbol{r}) \Delta\epsilon(\boldsymbol{r}) G(\boldsymbol{r}; \boldsymbol{r}_{\rm p}) \mathrm{d}^3\boldsymbol{r}. \tag{3.70}$$

$\mathcal{V}$ denotes the volume of the hologram, $\mathcal{E}_i$, the incident field, $\Delta\epsilon(\boldsymbol{r})$, the phase hologram, and $G(\boldsymbol{r}; \boldsymbol{r}_{\rm p})$ is the scalar Green's function for free space [84]:

$$\begin{aligned} G(\boldsymbol{r}; \boldsymbol{r}_{\rm p}) &= \frac{1}{i\lambda\,|\boldsymbol{r} - \boldsymbol{r}_{\rm p}|} \exp\left\{ i2\pi \frac{|\boldsymbol{r} - \boldsymbol{r}_{\rm p}|}{\lambda} \right\} \\ &\approx \frac{1}{i\lambda(z_{\rm p} - z)} \exp\left\{ i2\pi \frac{z_{\rm p} - z}{\lambda} + i\pi \frac{(x_{\rm p} - x)^2 + (y_{\rm p} - y)^2}{\lambda(z_{\rm p} - z)} \right\}. \end{aligned} \tag{3.71}$$

The last relation follows by expressing a spherical wave in the paraxial approximation. We assumed that $z_{\rm p} > z$ for all pairs of integration–observation points $z_{\rm p}$, z.

Similarly, the spherical reference wave (upon recording) is expressed in the paraxial approximation as in (3.60), repeated here for convenience,

$$R(\boldsymbol{r}) = \frac{1}{i\lambda(z + z_0)} \exp\left\{ i2\pi \frac{z + z_0}{\lambda} \right\} \exp\left\{ i\pi \frac{x^2 + y^2}{\lambda(z + z_0)} \right\}, \tag{3.72}$$

and the propagating signal is expressed as

$$S(\boldsymbol{r}) = \exp\left\{ i2\pi u_S \frac{x}{\lambda} \right\} \exp\left\{ i2\pi \left(1 - \frac{u_S^2}{2} \right) \frac{z}{\lambda} \right\}. \tag{3.73}$$

Then we have

$$\Delta\epsilon(\boldsymbol{r}) = R^*(x, y, z) S(x, z), \tag{3.74}$$

$$\mathcal{E}_i(\boldsymbol{r}) = R(x - \delta, y, z). \tag{3.75}$$

We will also assume that the spherical wave and the recording material are infinite in the transverse (x, y) directions, and that the thickness of the hologram is L in the

z direction. Substituting into (3.70), we obtain the following expression:

$$\mathcal{E}_d(\boldsymbol{r}_\mathrm{p}) \approx \frac{\exp\left\{i2\pi\frac{z_\mathrm{p}}{\lambda}\right\}}{\lambda^2 z_0^2}\int_{-\infty}^{+\infty}\mathrm{d}x\int_{-\infty}^{+\infty}\mathrm{d}y\int_{-\infty}^{+\infty}\mathrm{d}z\ \mathrm{rect}\left(\frac{z}{L}\right)$$

$$\exp\left\{-i\pi\left(u_S^2\frac{z}{\lambda}-\frac{\delta^2}{\lambda(z+z_0)}\right)\right\}\exp\left\{i2\pi\frac{x}{\lambda}\left(u_S-\frac{\delta}{z_0+z}\right)\right\}$$

$$\frac{1}{\lambda(z_\mathrm{p}-z)}\exp\left\{i2\pi\frac{z_\mathrm{p}-z}{\lambda}\right\}\exp\left\{i\pi\frac{(x_\mathrm{p}-x)^2+(y_\mathrm{p}-y)^2}{\lambda(z_\mathrm{p}-z)}\right\}. \quad (3.76)$$

The volume integral is calculated analytically as follows: The x and y integrals are readily obtained by using the following lemma from complex analysis [85]:

$$\int_{-\infty}^{+\infty}\exp\left\{i(aw^2+2bw)\right\}\mathrm{d}w = \sqrt{\frac{\pi}{|a|}}\exp\left\{i\left(\mathrm{sgn}(a)\frac{\pi}{4}-\frac{b^2}{a}\right)\right\}, \quad (3.77)$$

for a, b real and $a \neq 0$. Then we expand the denominators of the form $(z+z_0)^m$, $(m = 1, 2)$ in the exponents, keeping terms of order (z/z_0) only. The resulting z integral yields:

$$\mathcal{E}_d(\boldsymbol{r}_\mathrm{p}) \approx \frac{\exp\left\{i\frac{2\pi}{\lambda}\left[u_S x_\mathrm{p}+\left(1-\frac{u_S^2}{2}\right)z_\mathrm{p}\right]+i\frac{2\pi}{\lambda}\frac{\delta(x_\mathrm{p}-u_S z_\mathrm{p})}{z_0}\right\}}{\lambda^2 z_0^2}\times$$
$$\mathrm{sinc}\left\{\frac{\delta L}{\lambda z_0}\left(u_S-\frac{x_\mathrm{p}-u_S z_\mathrm{p}}{z_0}\right)\right\}. \quad (3.78)$$

The first term in (3.78) is explained as follows: if $\delta = 0$, the diffracted far field is a plane wave propagating in the direction u_S of the original signal. For $\delta \neq 0$, the direction of the reconstruction deviates by δ/z_0 (paraxially) from u_S. The direction-dependent sinc term suppresses the diffracted power, a result of phase-mismatch among wavelets produced in different positions along the volume hologram (Bragg mismatch). In the far field we can make the stationary phase assumption (i.e., assume that significant diffraction is obtained only at $x_\mathrm{p} \approx u_S z_\mathrm{p}$) to obtain for the diffraction

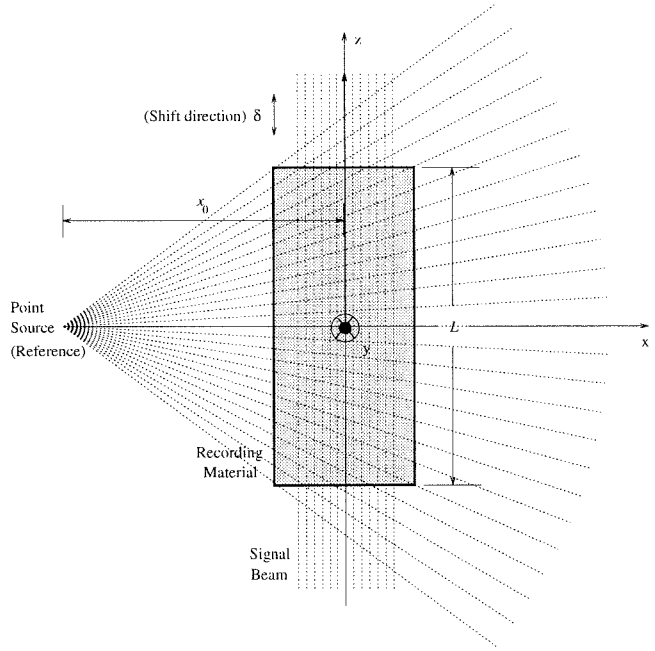


Figure 3.14: Geometry for shift multiplexing in the 90° geometry using a spherical reference wave.

efficiency

$$\eta(\delta) \equiv \frac{|\mathcal{E}_d|^2}{|\mathcal{E}_i|^2} \sim \text{sinc}^2\left(\frac{\delta u_S L}{\lambda z_0}\right). \tag{3.79}$$

Therefore, under the above assumptions, the Bragg nulls in diffraction efficiency occur at

$$\delta = m\delta_{\text{Bragg}} \equiv m\frac{\lambda z_0}{L u_S} \qquad m = 1, 2, \ldots \tag{3.80}$$

3.3.2 90-degree geometry

In the context of angle multiplexing, 90° (90-degree) geometry refers to the arrangement where the reference and signal beam during recording are perpendicular to each other and illuminate adjacent crystal faces [86]. Here we apply the idea of using a spherical reference beam to a similar arrangement, as shown in Fig. 3.14.

The reference is a spherical wave propagating along the x-axis, expressed as

$$R(x,y,z) = \frac{1}{i\lambda(x+x_0)} \exp\left\{i2\pi\frac{x+x_0}{\lambda}\right\} \exp\left\{i\pi\frac{z^2+y^2}{\lambda(x+x_0)}\right\}. \tag{3.81}$$

The signal is a plane wave propagating along the z axis,

$$S(z) = \exp\left\{i2\pi\frac{z}{\lambda}\right\}. \tag{3.82}$$

As in the case of transmission geometry, we perform shift multiplexing by translating the hologram by δ relative to the reference beam, except in the 90° geometry the translation must happen *along the direction of the signal beam*, i.e., in the z direction. A straightforward calculation along the lines of the Appendix of section 3.3.1 yields the following result for the diffraction efficiency as function of δ,

$$\eta(\delta) = \mathrm{sinc}^2\left(\frac{2L\delta}{\lambda x_0}\right). \tag{3.83}$$

Therefore, the shift selectivity in the 90° geometry is

$$\delta_m = m\frac{\lambda x_0}{2L} + \frac{\lambda}{2(\mathrm{NA})} \qquad m = 1, \ldots, \tag{3.84}$$

where the second component was added to account for the finite extent of the point source, as in the case of transmission geometry.

3.3.3 Fractal shift multiplexing

So far we assumed that shift multiplexing is performed by translating the reference relative to the hologram in the plane defined by the optical axis of the reference and the optical axis of the signal. In the previous two sections we showed that the translation results in Bragg mismatch. This effect allows us to multiplex holograms by translating by integer multiples of the Bragg selectivity between successive exposures. One wonders if something similar happens when the reference is translated with respect to the hologram in the perpendicular y-direction, as shown in Fig. 3.15.

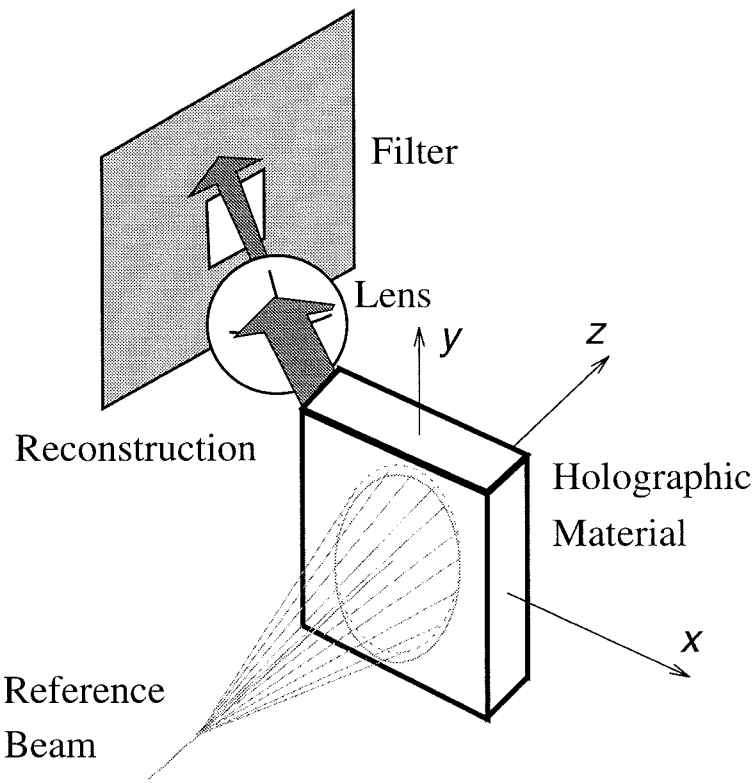


Figure 3.15: Geometry for fractal shift multiplexing.

The answer in that case is that the Bragg shift selectivity is very poor, because of the "Bragg-degeneracy" occurring in the perpendicular direction, an effect similar to the degeneracy in out-of-plane angle multiplexing (see section 3.1 – Appendix II). The expression for the Bragg selectivity is calculated with the same method used in section 3.3.1 for the in-plane selectivity. The result is

$$\delta_y = z_0 \sqrt{\frac{2\lambda}{L}} \tag{3.85}$$

Comparing (3.67) and (3.85) we observe that the out-of-plane selectivity may be orders of magnitude larger than the Bragg selectivity; therefore, shift multiplexing in the orthogonal (y) direction is much sparser than in the regular (x) in-plane case. In most situations, we can increase the number of holograms that can be superimposed in the y-direction by using the property of spherical-reference holograms to rotate

when the recontructing reference is translated in any direction. In section 3.3.1 we saw that this rotation produces Bragg mismatch when the reference translates in the x direction. If the reference translation is in the y direction, we can use this property to multiplex holograms as shown in Fig. 3.15, by inserting to the reconstruction path a lens and an aperture in the Fourier plane. The lens transforms rotation to translation, and the reconstruction is blocked completely when the rotation is enough to move the reconstructed spectrum out of the aperture. A new hologram may be recorded in the translated position, because the contribution from the previous hologram is minimal (corresponding only to the higher-order diffraction lobes). This method of shift multiplexing is reminiscent of the fractal and peristrophic methods when the reference is a plane wave, and we will call it "fractal shift multiplexing." The fractal shift selectivity, i.e., the required amount of y-translation before the reconstructed spectrum is blocked by the aperture, is given by

$$\delta_y{}' = \begin{cases} N_p b z_0 / F & \text{for Fourier plane holograms} \\ 2\lambda z_0 / b & \text{for image plane holograms} \end{cases} , \tag{3.86}$$

where N_p is the number of pixels and b is the pixel size of the stored image.

Usually $\delta_y{}' < \delta_y$; therefore, the fractal method is preferrable to the Bragg selectivity method in the y direction. If $\delta_y{}'$ and δ_y are comparable, the fractal method is still preferrable because usually there is strong crosstalk between holograms multiplexed in the out-of-plane direction [87]. The reason is that the motion of holograms on the k-sphere in response to y translation is such that Bragg mismatch does not occur simultaneously to the entire image[3]. The Bragg method is preferrable only if $\delta_y << \delta_y{}'$, which, however, occurs only rarely in practical situations. Experimentally, the fractal method was used to increase the capacity of shift multiplexed holographic disks by a factor of approximately 2 in [33, 87].

[3]This is not entirely true even for x translation; see section 4.1. However, the crosstalk in that case is first order, i.e., at least one order of magnitude smaller than the reconstruction. In the case of fractals, crosstalk is of the same order of magnitude as the reconstruction [88] (zero-th order!).

Chapter 4 Shift multiplexed storage systems

In this chapter we concentrate on the implementation of shift multiplexing using a spherical wave reference. In section 4.1 we derive theoretically and present experimental results on crosstalk between holograms superimposed using the shift multiplexing method, and show that crosstalk behaves approximately the same as in the case of angle multiplexed holograms. In section 4.2 we address the issue of dynamic range for shift-multiplexed memories in photorefractive materials, and give two alternative exposure schedules, sequential and interleaved recording. We demonstrated the sequential technique by storing 600 holograms in lithium niobate. A peculiar effect of sequential recording on the spectral properties of shift-multiplexed holograms is described in section 4.3.

The storage density of angle and wavelength multiplexed holographic 3–D disks was derived in [44]. It was shown that uniformity considerations for the edges of the stored holograms cause the density to peak at a theoretical maximum of 117.2bits/μm^2 (for typical SLM parameters and optical apertures) for a 16.7 mm thick lithium niobate disk, using four symmetric reference angles for recording. In section 4.4 we present the corresponding derivation for shift multiplexing. We show that the density of a shift multiplexed disk increases monotonically with thickness, and eventually saturates. Finally, in section 4.5 we analyze diffraction from a volume hologram recorded with a spherical reference when the readout wavelength is other than the recording wavelength.

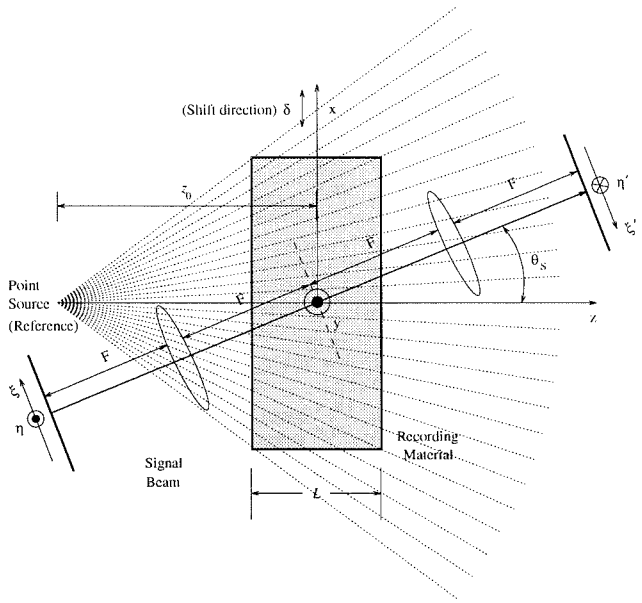


Figure 4.1: Geometry for the theoretical calculation of crosstalk in shift multiplexing using a spherical reference wave.

4.1 Cross-talk in shift-multiplexed holographic memories

The approximate theory presented in the Appendix to section 3.3.1 predicts that the diffraction efficiency η of spherical volume holograms as function of shift has nulls at integer multiples of δ_{Bragg}. This holds for the ideal situation of a hologram that is infinite in the transverse directions, recorded using a spherical wave of zero spot size as reference and a plane wave as signal. We also neglected the variable modulation depth effects due to the variation in intensity of the reference and the signal throughout the volume of the hologram. Finally, the calculation was performed for a single signal component incident at θ_S. In general, the signal occupies a finite–size bandwidth in reciprocal space, and hence each component Bragg-mismatches at different δ.

In this section, we develop a theoretical model for the crosstalk induced by the finite signal bandwidth in the case of shift multiplexing in the Fourier plane. In the calculation we will drop the dependence of the selectivity on numerical aperture (i.e.,

the Δx correction). The assumption of an infinite spherical wave for shift multiplexing is equivalent to assuming infinite plane wave reference for other methods, as was done in calculations of crosstalk for angle [71], wavelength [73, 72] and phase code [75] multiplexing in the Fourier plane, and for image plane holograms [89]. We show that, under these assumptions, the results for shift multiplexing are consistent with the angle multiplexing analysis. Then we characterize the crosstalk experimentally and compare the results with the theory.

Consider the Fourier plane geometry of Fig. 4.1. Let $f_m(\xi,\eta)$, $m = 1,\ldots,M$ denote the pattern stored as the m-th page. M is the maximum number of overlapping pages on any location. The signal corresponding to the m-th hologram is expressed as:

$$\begin{aligned} S_m(x,y,z) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \mathrm{d}\xi\mathrm{d}\eta f_m(\xi,\eta) \\ &\exp\left\{-i2\pi\frac{\xi}{\lambda F}\left(-\sin\theta_S z+\cos\theta_S x\right)-i2\pi\frac{\eta y}{\lambda F}\right\} \\ &\exp\left\{i\frac{2\pi}{\lambda}\left(1-\frac{\xi^2+\eta^2}{2F^2}\right)\left(\cos\theta_S z+\sin\theta_S x\right)\right\}. \end{aligned} \tag{4.1}$$

In order to reconstruct hologram m', the recording area is illuminated by a spherical beam displaced by $m'\delta$:

$$\mathcal{E}_i(m') = \frac{1}{i\lambda(z+z_0)}\exp\left\{i2\pi\frac{z+z_0}{\lambda}\right\}\exp\left\{i\pi\frac{(x-m'\delta)^2+y^2}{\lambda(z+z_0)}\right\}. \tag{4.2}$$

The diffracted field is obtained using the theory of Appendix I with the paraxial approximation $\sin\theta_S \equiv u_S << 1$, $\cos\theta_S \approx 1-u_S^2/2$, and neglecting refraction. A lengthy but straightforward calculation yields for the detector plane the following expression:

$$\begin{aligned} &\mathcal{E}_{m'}(\xi',\eta') \approx \\ &\sum_{m=0}^{M-1} f_m\left(\xi'+\frac{(m-m')\delta}{z_0}F,\eta'\right)\mathrm{sinc}\left\{\frac{(m-m')\delta L}{\lambda z_0}\left(u_S-\frac{\xi'}{F}\right)\right\}. \end{aligned} \tag{4.3}$$

A straightforward calculation along the lines of [71] shows that a similar expression holds approximately for the crosstalk in the geometry of Fig. 4.1 (in the paraxial approximation) if we replace the spherical reference wave with a plane wave parallel to the z axis and perform angle multiplexing instead of shift multiplexing. A significant difference between the cases of shift and the exact solution for angle multiplexing is that, in the former, symmetry makes crosstalk depend on the difference $m - m'$ only.

When reconstructing hologram m', the fact that the remaining multiplexed holograms were recorded displaced by a multiple of the shift selectivity δ guarantees only that their central component, i.e., the central pixel $\xi' = 0$, will be Bragg mismatched. All other locations in the multiplexed images still diffract weakly, because their shift selectivity is given by (3.64) with $u_S - \xi'/F$ rather than u_S. These contributions appear as crosstalk around the noise-free central pixel.

Let us assume that a large number M of Fourier plane holograms are shift multiplexed, and are separated by p shift Bragg nulls, i.e., the relative translation between successive recordings is $p\delta$, where δ is the shift selectivity. Under the image statistics assumed in [71, 72, 75], the expected value of the crosstalk noise power is given by the expression

$$P_{XN} \approx \sum_{m=0}^{M-1} \mathrm{sinc}^2 \left\{ p(m - m') \left(1 - \frac{\xi'}{u_S F} \right) \right\}, \tag{4.4}$$

where the signal power was taken equal to 1. If

$$\frac{p\,|\xi'|}{u_S F} << 1 \qquad \text{and} \qquad M \to \infty, \tag{4.5}$$

then the summation can be carried out analytically, and yields

$$P_{XN} \approx \frac{|\xi'|}{2pu_S F}. \tag{4.6}$$

Therefore, at pixels lying close to the carrier u_S, the noise increases linearly with distance from the image center, and is inversely proportional to the null order p.

Theoretical plots of the exact relation (4.4) are given in Fig. 4.2. As the pixel

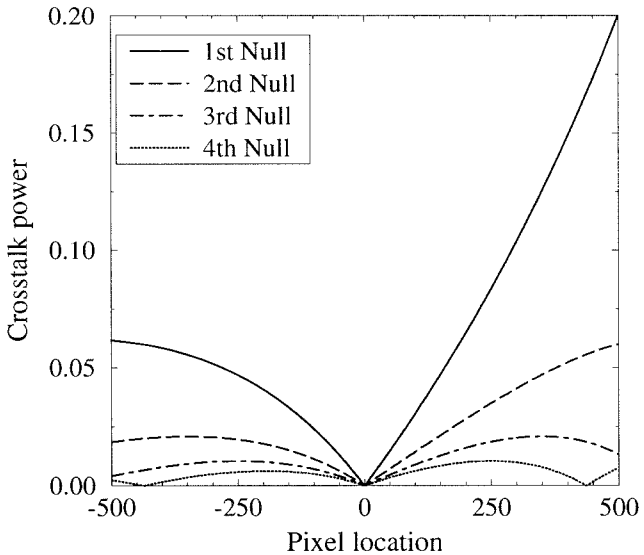


Figure 4.2: Theoretical plots of expected crosstalk power versus pixel location for Fourier plane shift multiplexed holograms. The parameters used for the plots were: hologram thickness $L = 1$ mm, angle of incidence of the signal $\theta_S = 20°$, wavelength $\lambda = 488$ nm, focal length $F = 5$ cm, pixel size $b = 10\mu$m.

value increases, (4.5) is violated, and the noise pattern becomes asymmetric. Pixels with large positive values are closer to the z-axis and suffer from higher noise. The same curves hold approximately for angle multiplexing in the off–axis geometry, if the same parameters (including the number of holograms M) are used.

In order to characterize the crosstalk effects for shift-multiplexed volume holograms recorded with spherical reference beams, we performed the following experiment: We stored 20 holograms of rotated versions of the same chessboard pattern in 21 shift multiplexed positions, leaving position #11 blank. Therefore, excess light measured in the location of hologram #11 is due to crosstalk contributions from the neighboring holograms. The shift separation between adjacent holograms was chosen equal to δ, $2\,\delta$, $3\,\delta$, and $4\,\delta$ (i.e., $p = 1, 2, 3, 4$, respectively), where for δ we used the experimentally determined value 3.7μm. The holograms were stored in the Fresnel region. The crosstalk theory developed for Fourier plane holograms also applies to Fresnel holograms recorded anywhere between the two lenses in a 4F imaging system.

Behavior consistent with that predicted in Fig. 4.2 is observed in Fig. 4.3, where the cross section of the reconstruction of location #11 is plotted (in the absence

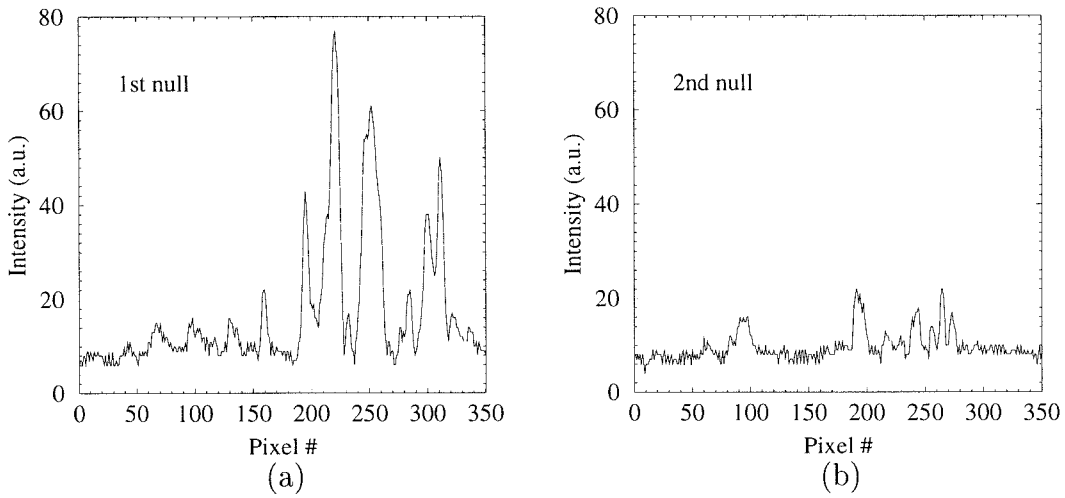


Figure 4.3: Cross sections of the diffracted pattern at shift location #11 (originally left blank) when the surrounding holograms are multiplexed (a) in the 1st Bragg null and (b) in the 2nd Bragg null. The units on both axes are arbitrary, but horizontal and vertical scales are the same in both plots.

of crosstalk, this would contain only scatter and detector noise contributions). The asymmetry predicted in (4.3) is evident for storage in the 1st null. In the case of using the 2nd null, the noise power decreases considerably and the asymmetry becomes less pronounced, in agreement with the theoretical curves of Fig. 4.2.

SNR results are given in Fig. 4.4 for the cases of a single hologram and 21 multiplexed holograms. In the case of a single hologram, we calculated the SNR by measuring the spatially averaged diffracted power from the hologram at zero translation, and dividing by the diffracted power at shifts equal to δ, $2\,\delta$, $3\,\delta$, and $4\,\delta$. For the multiple holograms, we calculated the SNR by dividing the diffraction efficiency at location #10 with the diffraction efficiency at location #11 (empty slot) for the four cases of null separation.

In the same plot we also give the theoretical values of the ratio between the expected total signal power and the expected total noise power, for each case of null separation. The three curves show the same qualitative behavior, although there is a noticeable discrepancy between the theoretical and experimental values of crosstalk–induced SNR. The saturating behavior of the experimental data indicates that the discrepancy is mainly due to noise sources unrelated to crosstalk, such as scatter

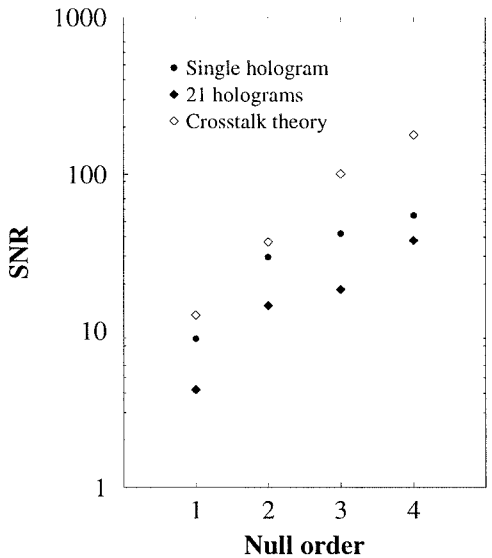


Figure 4.4: Signal–to–Noise ratio (SNR) versus null order p (in multiples of $\delta = 3.7$ μm) for two experiments: single hologram and 21 holograms. Shown also is the theoretical SNR prediction for the maximum number M of allowable shift multiplexed holograms at the respective null orders.

noise and multiple reflections off the uncoated crystal surfaces. In addition, small contributions are present from crosstalk sources such as finite numerical aperture and variable modulation depth that we neglected in the theory.

4.2 Exposure schedule and dynamic range issues

The diffraction efficiency η of holograms recorded in diffusion dominated photorefractive materials is described as a function of the recording time t by a saturating exponential of the form (see, e.g., [35, 90]):

$$\eta(t) = \eta_0 \left(1 - \exp\left\{-t/\tau_{\mathrm{w}}\right\}\right)^2, \tag{4.7}$$

where η_0 is the saturation diffraction efficiency and τ_{w} is the recording time constant. On the other hand, when a hologram of strength η_1 is illuminated, it decays exponentially as

$$\eta(t) = \eta_1 \, \exp\left\{-2t/\tau_{\mathrm{e}}\right\}, \tag{4.8}$$

where τ_e is the erasure time constant. The parameters τ_w, τ_e depend on the geometry, the total exposure power, the modulation depth and the absorption coefficient. A detailed calculation of the time constants is outside the scope of this thesis. Instead, we will assume that the exponential models of (4.7) and (4.8) hold, and we will determine the value of τ_e experimentally.

In multiplexing techniques based on recording over the same spot (e.g., angle multiplexing), holograms recorded early are erased by their successors. The first holograms are erased more, thus they must be initially stronger; this requirement was used in [51, 91, 92] to derive hologram recording times as function of hologram order. This function is referred to as "exposure schedule." The exposure times depend on τ_w, τ_e and the number of holograms M.

We will now describe the "sequential" recording exposure schedule for shift multiplexed holograms. With this method, shift-multiplexed holograms are recorded by rotating the disk by an angle ϕ_{disk} sufficient to produce translation equal to the shift selectivity δ between successive exposures. Let R be the radius of the track being recorded. Then ϕ_{disk} is given by

$$\phi_{\text{disk}} = \frac{\delta}{R} = \frac{\lambda z_0}{RL \tan\theta_S} + \frac{\lambda}{2R(\text{NA})}. \tag{4.9}$$

Fig. 4.5(a) shows how the sequential exposure schedule evolves in time and space. M is the number of shift multiplexed holograms that overlap within one spot. It is equal to the hologram aperture along the shift direction divided by δ (see section 4.4 for a derivation). The total number of holograms fitting in the track is given by

$$N = \frac{2\pi R}{\delta}. \tag{4.10}$$

A hologram is erased by its neighbors that start to its right and overlap vertically in the plot; thus A_m is erased by A_{m+1}, ..., A_{m+M-1} but not by the subsequent holograms. This is true for all indices m running from $M+1$ to $N-M$. Holograms A_1, ..., A_M will be further erased by holograms A_{N-M+1}, ..., A_N when the disk

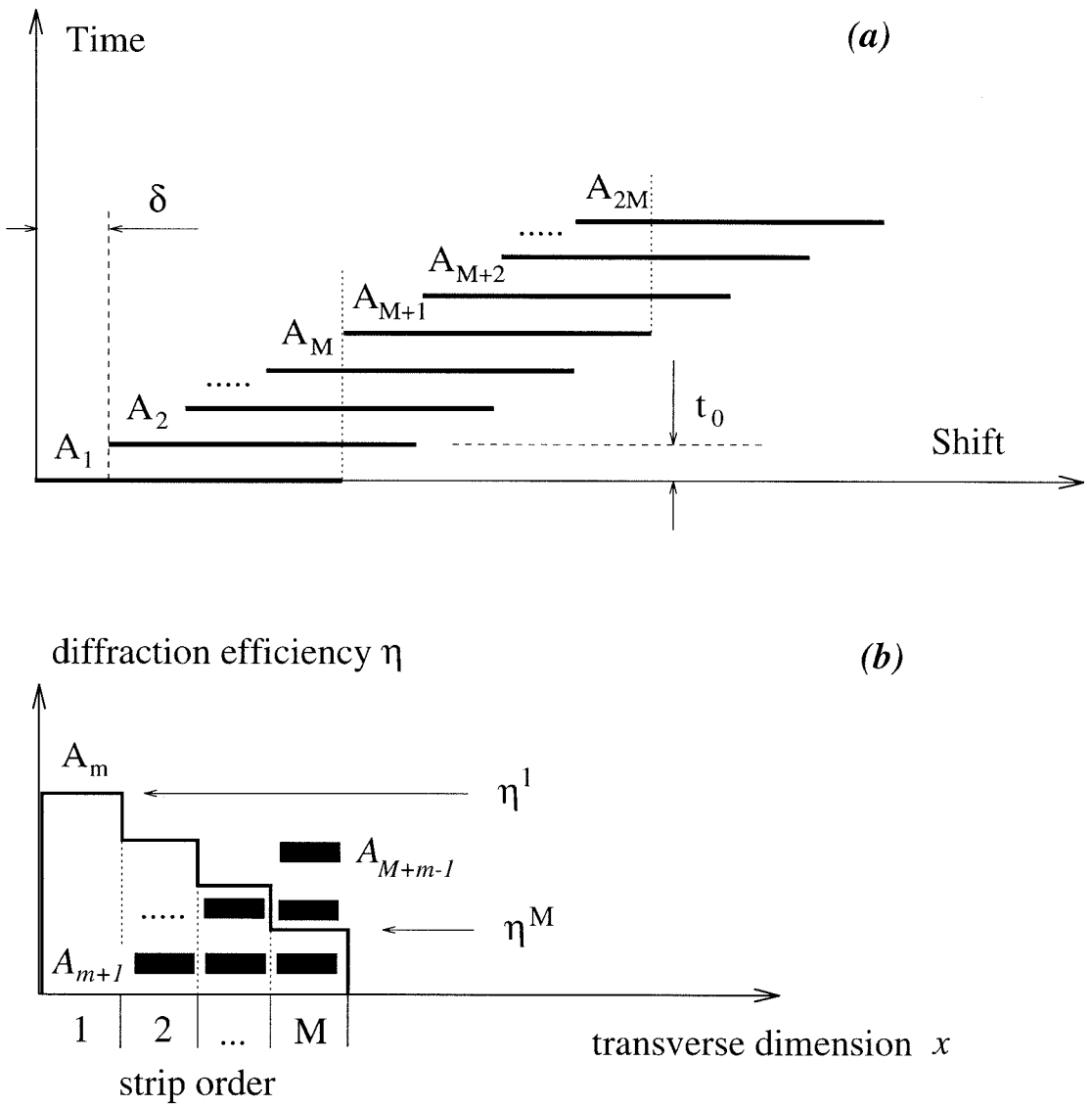


Figure 4.5: *(a)* Exposure schedule for sequential recording. Horizontal axis is shift, vertical is recording time. Bars A_1, A_2, ... denote holograms; the index corresponds to location on the disk; the horizontal location of a hologram in the graph denotes its shift with respect to the origin (left edge of the first hologram A_1), and the vertical location, the beginning of its exposure in the schedule. The horizontal separation is equal to the shift selectivity δ; the vertical separation is equal to the constant exposure time t_0 (see text). *(b)* Non-uniform erasure of hologram A_m by its successors A_{m+1}, ..., A_{m+M-1}. The diffraction efficiency curve follows the profile of A_m after recording of all its shift multiplexed neighbors is complete (see also section 4.3 and Figure 4.8).

completes one full revolution, whereas holograms A_{N-M+1}, ..., A_N will be erased less than the other holograms. Neglecting these edge effects, all other holograms are erased in the same manner, hence their diffraction efficiencies are equalized if they are recorded with the same exposure time t_0.

A consequence of the sequential approach is transverse non–uniformity as shown in Fig. 4.5(b). Consider any hologram A_m, except for the first M and the last M. The diffraction efficiency of A_m immediately after recording is given by

$$\eta^1 = \eta_0 \left(1 - \exp\left\{-t_0/\tau_{\rm w}\right\}\right)^2 . \tag{4.11}$$

The next hologram in the sequential schedule is A_{m+1} and it is recorded after shifting by δ. Thus it will erase A_m for a time t_0, except for a strip of width δ, which is denoted as strip 1 in Fig. 4.5(b); this strip will retain diffraction efficiency η^1. In general, after the end of the recording process, strip l of any hologram will have reached diffraction efficiency

$$\eta^l = \eta_0 \left(1 - \exp\left\{-t_0/\tau_{\rm w}\right\}\right)^2 \exp\left\{-2(l-1)t_0/\tau_{\rm e}\right\} . \tag{4.12}$$

Maximum erasure is suffered by strip M. The diffraction efficiency η^M of this strip is maximized if we choose

$$t_0 = \tau_{\rm w} \ln\left(1 + \frac{\tau_{\rm e}}{(M-1)\tau_{\rm w}}\right) \approx \frac{\tau_{\rm e}}{M}, \tag{4.13}$$

and is given by

$$\eta^M = \eta_0 \left(\frac{\tau_{\rm e}}{(M-1)\tau_{\rm w}}\right)^2 \left[1 + \frac{\tau_{\rm e}}{(M-1)\tau_{\rm w}}\right]^{-2M\frac{\tau_{\rm w}}{\tau_{\rm e}}} \approx \eta_0 \frac{\tau_{\rm e}^2 {\rm e}^{-2}}{M^2 \tau_{\rm w}^2}. \tag{4.14}$$

The approximations hold for $M >> 1$. We now define the average diffraction efficiency of the non-uniform holograms as:

$$\eta_{\rm av} = \frac{\int \eta(x) {\rm d}x}{\int {\rm d}x}, \tag{4.15}$$

where the integrals are along the aperture of the holograms in the shift direction. This results in

$$\eta_{\rm av} = 4\frac{\eta_0}{M}\exp\left\{-t_0\left(\frac{1}{\tau_{\rm w}}+\frac{M-1}{\tau_{\rm e}}\right)\right\}\frac{\sinh^2\dfrac{t_0}{2\tau_{\rm w}}\sinh M\dfrac{t_0}{\tau_{\rm e}}}{\sinh\dfrac{t_0}{\tau_{\rm e}}} \tag{4.16}$$

$$\approx \eta_0\,\frac{\tau_{\rm e}^2\left(1-{\rm e}^{-2}\right)}{2M^2\tau_{\rm w}^2}. \tag{4.17}$$

Eq. (4.17) results from (4.16) if we substitute the optimal value of t_0 calculated in (4.13). Thus, in the sequential schedule the average diffraction efficiency follows the $1/M^2$ rule but it is actually weaker than the diffraction efficiency of angle multiplexed holograms by a factor of $(1-{\rm e}^{-2})/2 \approx 0.432$. On the other hand, from (4.16) we observe that if we let $t_0 \to \infty$, $\eta_{\rm av}$ behaves like $1/M$. This expresses the fact that if we overexpose the holograms in the sequential method, then only the first strip of each hologram will survive, and the rest of the hologram will be erased. This situation is undesirable, since it restricts the recording area to a strip of width δ only, and degenerates shift multiplexing to spatial multiplexing, resulting in severe losses in storage density.

At the leading edge, holograms $\rm A_1$, ..., $\rm A_M$ have uniform diffraction efficiency equal to η^M, because they receive additional exposure at the end of the schedule, when the disk is about to complete one revolution. At the trailing edge, the worst affected strip of hologram $\rm A_{N-m}$, $m = M-1, \ldots, 1$, is $l = m+1$, and has diffraction efficiency $\eta^l \approx \eta_0\tau_{\rm e}^2{\rm e}^{-2(l-1)/M}/\tau_{\rm w}^2M^2$. Hologram $\rm A_N$ is uniform, since it is never erased, and has diffraction efficiency η^1.

The non–uniformity can be cancelled in Image plane holograms by recording with the inverse intensity dependence. Alternatively, for digital storage, one can record uniform holograms and use variable decision thresholds. Either method will yield good results if the diffraction efficiency of the most affected areas is kept sufficiently strong compared to the noise level by using the optimal t_0 of (4.13). On the other hand, the

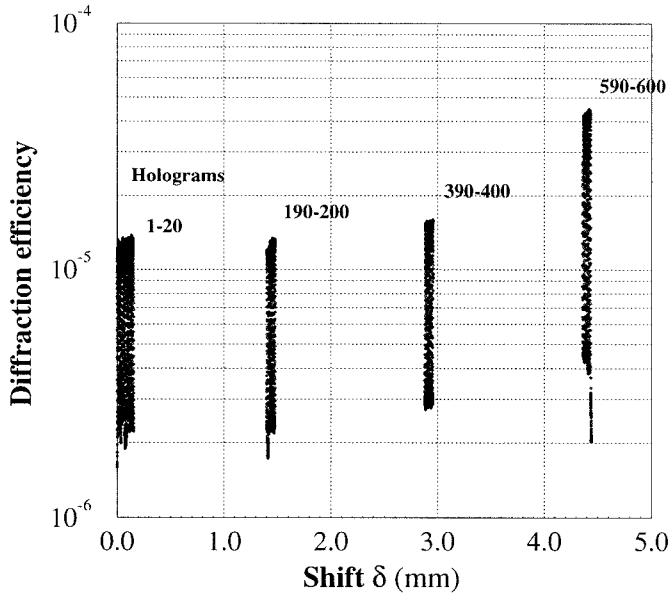


Figure 4.6: Plot of measured diffraction efficiency (after spatial integration by a single detector) of 50 out of 600 holograms stored with the sequential method. For the shift separation $\delta_{\text{shift}} = 7.4\mu$m (second null), and aperture size $s \approx 3$ mm, we have $M \approx 400$. Therefore only the first 200 holograms received equal exposure. The exposure time used in this experiment was $t_0 = 10$ sec.

non–uniformity has severe effects on Fourier holograms, since it shapes the hologram spectrum asymmetrically. This non–uniform filtering effect causes pixel broadening (intra–page noise); therefore, the contrast ratio of the reconstruction decreases with respect to the unfiltered case. For holograms recorded in the Fresnel region, image and Fourier plane effects are combined in the sense that one observes non–uniformity across the reconstruction and also a decrease in the contrast ratio. In section 4.3 we characterize theoretically the erasure induced by non–uniform filtering.

We used the sequential exposure schedule to record 600 holograms in the experimental set–up of Fig. 3.12. We set the separation between adjacent holograms to 7.4 μm, which equals twice the measured shift selectivity $\delta = 3.7\ \mu$m. The size of the signal beam projected onto the crystal surface was approximately 3 mm. Therefore, the number of overlapping holograms in this experiment was $M \approx 400$. For the crystal we used and the given geometry, we measured $\tau_e \approx 3,500$ sec. For recording we used t_0=10 sec as the constant exposure time. Each reconstruction was spatially

integrated onto a single detector in order to measure the diffraction efficiency. The results are plotted in Fig. 4.6. It is seen that the first 200 holograms were successfully equalized in terms of the total diffraction efficiency, as they all received equal exposure. From then on, diffraction efficiency vs. hologram number attains an upward slope, as expected, since as the order of holograms increases, the number of overlapping holograms decreases.

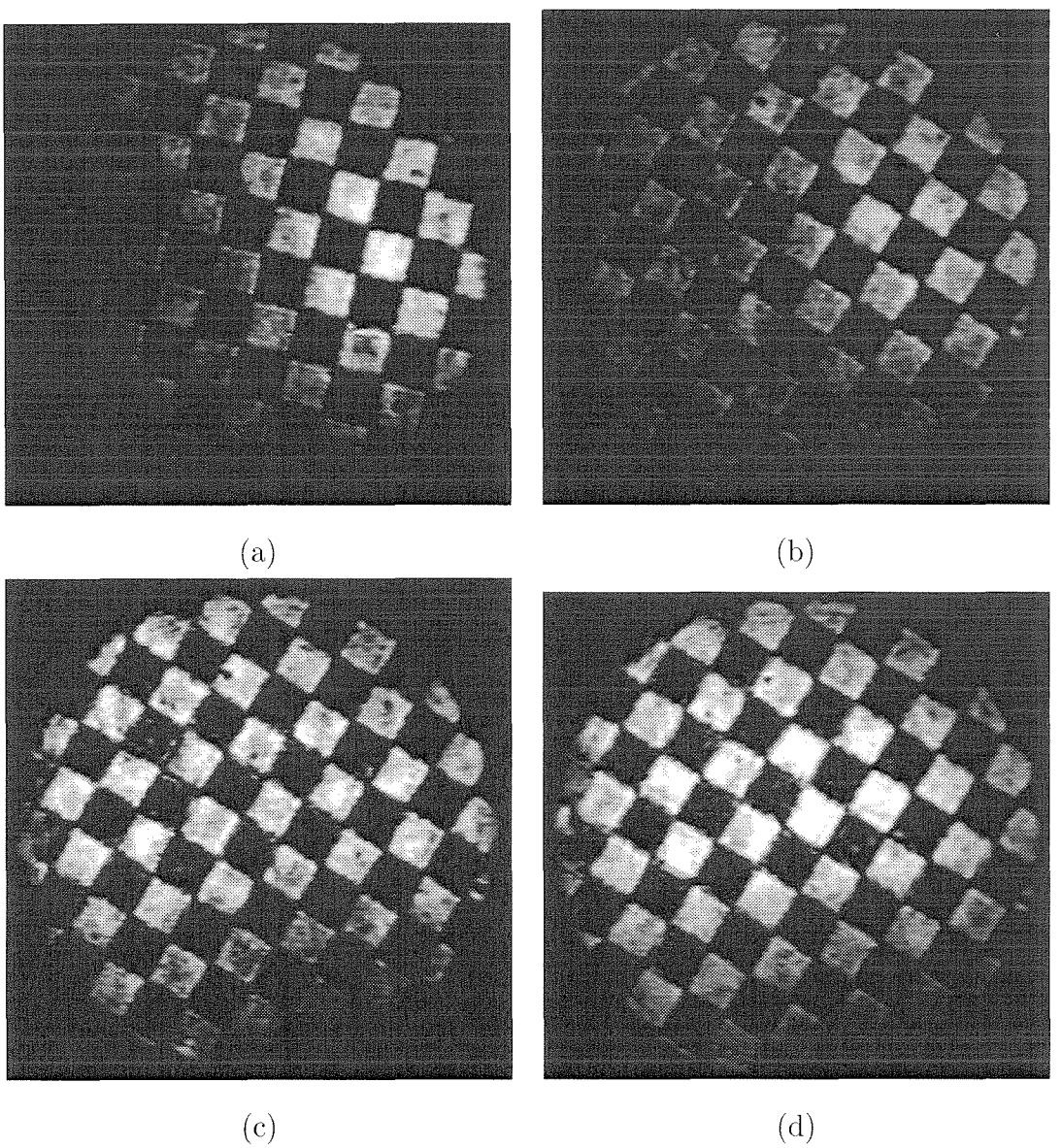


Figure 4.7: Reconstructions of holograms (a) 1, (b) 200, (c) 400, (d) 600 from the experiment of Figure 4.6. Shift direction was from left to right.

In Fig. 4.7 we show a few reconstructions from the 600 holograms. All holograms (with the exception of the last few) exhibit non–uniformity towards the shift direction. Since the image features were quite large in this experiment, the pixel broadening effect was not observed.

We can eliminate the non–uniformity through the use of a different, "interleaved," exposure schedule. In this scheme, we record one complete track of non–overlapping (spatially multiplexed) holograms before moving to the next shift multiplexed position. This method is well matched to the disk configuration since we can record a new set of slightly shifted spatially multiplexed holograms during each disk rotation. Interleaving works perfectly if $N+1$ is an integer multiple of M, otherwise the first M and last M holograms suffer from over– and under–exposure respectively, as in the sequential method. These small edge effects can be ignored in practice.

As described, recording consists of M epochs. At epoch q ($q = 0, \ldots, M-1$) we record holograms A_q, A_{M+q}, $\ldots$, A_{N-M+q}. The recording time for all holograms at epoch q is t_q. Because full tracks are recorded so that they completely overlap (but still they are displaced by δ with respect to each other), all holograms are erased uniformly; moreover, tracks recorded later are erased less than their predecessors. The uniform diffraction efficiency of the holograms after recording epoch q is given by

$$\eta(q) = \eta_0 \left(1 - \exp\left\{-\frac{t_q}{\tau_{\mathrm{w}}}\right\}\right)^2 \exp\left\{\sum_{q'=q+1}^{M-1} \frac{2t_{q'}}{\tau_{\mathrm{w}}}\right\}. \tag{4.18}$$

This same equation holds for methods of complete overlap, e.g., angle multiplexing [91]. Therefore, the exposure schedule is determined identically. The optimal diffraction efficiency is given by

$$\eta \approx \eta_0 \frac{\tau_{\mathrm{e}}^2}{M^2 \tau_{\mathrm{w}}^2}. \tag{4.19}$$

It is the same for all holograms, and equal to the diffraction efficiency yielded by the exposure schedule for angle multiplexed holograms. The price to pay for the

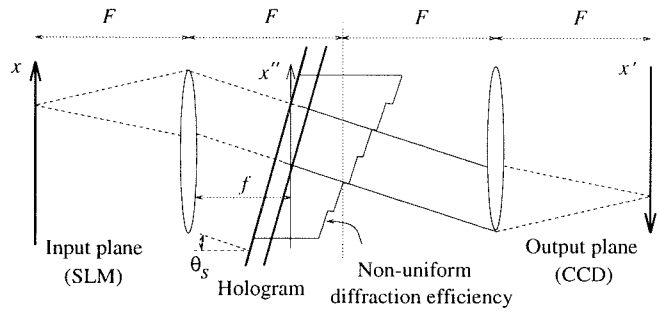


Figure 4.8: Geometry for the calculation of the distortion occurring in shift multiplexed holograms recorded in photorefractive materials, due to partial erasure in the Fourier or Fresnel regions. The filter is shift variant if the hologram is not centered with respect to the Fourier plane. See also Figure 4.5.

equalization provided by the interleaving method is considerable complication in the recording process.

4.3 Distortion due to non-uniform erasure

Shift–multiplexed holograms in photorefractive crystals stored using the sequential schedule suffer non–uniform erasure (see Section 4.2). We will now characterize this effect for holograms stored in the Fresnel region, and in the Fourier plane as a special case.

The geometry used for the calculation is shown in Figure 4.8. The hologram is tilted with respect to the signal beam path by angle θ_S, and is located distance f from the Fourier transforming lens (focal length F). For simplicity we will ignore the thickness of the recording material, and the possible aberrations introduced by the tilted path. The shift selectivity is δ, and the pixel size is b. We assume that during recording, the signal is low–pass filtered at the Nyquist cut–off frequency $2\lambda F/b$ so that the area it takes on the disk is minimized without any loss in information content. Because of the shift multiplexing mechanism, successive slices of the hologram suffer exponential erasure by an amount t_0 compared to their neighbors. Thus the diffraction

efficiency is given by the staircase–like function:

$$\eta(x'',x) = \sum_{l=0}^{m_0-1} \exp\left\{-i2\pi\frac{xx''}{\lambda F}\right\} \exp\left\{-(l_0+l)t_0/\tau_e\right\} \times \operatorname{rect}\left\{\frac{x'' - \left(l - \frac{m_0-1}{2}\right)\beta\delta}{\beta\delta}\right\}, \tag{4.20}$$

where the indices l_0 and m_0 indicate which part of the staircase corresponds to the point source located at x, and β is a correction factor for the tilt. These parameters are obtained directly from the geometry of Figure 4.8, and are given by the following expressions:

$$l_0(x,f) = \left|F - f - \frac{\lambda F}{b}\tan\theta_S\right| \times \begin{cases} \dfrac{\frac{N_p b}{2} - x}{F\delta\cos\theta_S}, & \text{if } f < F - \frac{\lambda F}{b}\tan\theta_S, \\ \dfrac{\frac{N_p b}{2} + x}{F\delta\cos\theta_S}, & \text{if } f > F - \frac{\lambda F}{b}\tan\theta_S, \end{cases} \tag{4.21}$$

$$\beta(x) = \cos\theta_S + \frac{x}{F}\sin\theta_S, \tag{4.22}$$

$$m_0(x) = \frac{2\lambda F}{b\beta(x)\delta}. \tag{4.23}$$

The total number of strips M is needed in order to determine the optimum recording time according to the theory of section 4.2. M depends on the defocusing distance f and is given by

$$M(f) = \begin{cases} \dfrac{2\lambda F}{b\delta\cos\theta_S}\left(1 + \left|1 - \dfrac{f}{F}\right|\dfrac{N_p b^2}{2\lambda F}\right), & \text{if } |F-f| > \frac{\lambda F}{b}\tan\theta_S, \\ \dfrac{2\lambda F}{b\delta\cos\theta_S}\left(1 + \dfrac{N_p b}{2F}\tan\theta_S\right), & \text{if } |F-f| < \frac{\lambda F}{b}\tan\theta_S. \end{cases} \tag{4.24}$$

The transfer function is then determined as:

$$h(x',x) = \mathrm{sinc}(\kappa)\,\mathrm{ar}\,(\kappa+i\zeta;m_0)\,\exp\left\{-2\pi\left(l_0+\frac{m_0-1}{2}\right)\zeta\right\}, \tag{4.25}$$

$$\kappa = \frac{(x'-x)\beta(x)\delta}{\lambda F}, \tag{4.26}$$

$$\zeta = \frac{t_0}{2\pi\tau_e}, \tag{4.27}$$

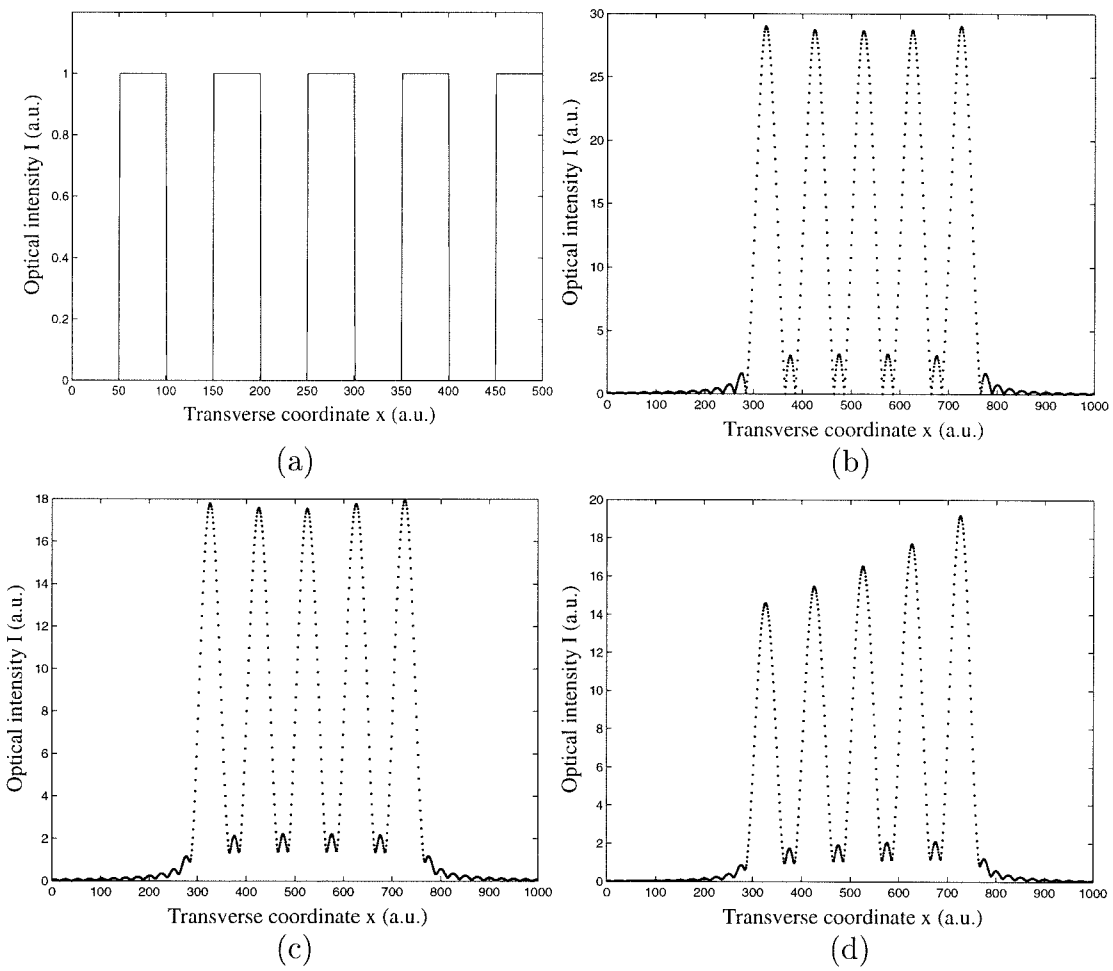


Figure 4.9: Effects of shift–induced non-uniformity on Fourier and Fresnel holograms. (a) Original chessboard pattern. (b) Nyquist filter (cut–off at $\pm\lambda F/b$) without absorption ($\tau_e = \infty$), located at $f = F = 5$ cm. (c) Nyquist filter with $t_0/\tau_e = 0.011$, $f = F = 5$ cm (Fourier filter). (d) Nyquist filter with $t_0/\tau_e = 0.0092$, $f = 4$ cm (Fresnel filter).

where the array function $\mathrm{ar}(u;l)$ was defined in eq. 3.55 in section 3.2. Note that the filter represented by this transfer function is shift variant (unless $f = F$ and $\theta_S = 0$, which would yield very bad shift selectivity). In general, the diffraction efficiency is asymmetric (except when $f = F$), in agreement with experiment (see Fig. 4.7). The weaker edge is towards the shift direction if $f < F$ and in the opposite direction otherwise. The resolution is worse than the case of no erasure ($\tau_e = \infty$) and decreases uniformly towards the weaker edges.

Some sample simulated reconstructions are shown in Fig. 4.9. The parameters used for this numerical example were $\lambda = 488$ nm, $F = 5$ cm, $N_p = 10$, $b = 100$ μm, $\theta_S = 40°$, $\delta = 7$ μm. The original pattern used for the simulations is shown in Fig. 4.9(a). In Fig. 4.9(b) we have plotted the reconstruction for $f = 4$ cm with no absorption ($t_0/\tau_e = 0$). In that case simple low–pass filtering takes place, with cut–off frequency equal to the Nyquist frequency $2\lambda F/b$ determined for intensity detection. The contrast ratio is $\mu = 91.41$ in this example. For a Fourier filter with $t_0/\tau_e = 0.011$ (approximately equal to $1/M$, where $M = 92$ for this case), Fig. 4.9(c), the contrast ratio drops to $\mu \approx 72$, 21.2% down with respect to the simple Nyquist filter. Finally, in Fig. 4.9(d) the result of a Fresnel filter with $f = 4$ cm, $M = 109$ and $t_0/\tau_e = 1/M = 0.0092$ is shown. The contrast ratio is $\mu = 72.15$ at the weak edge and $\mu = 83.04$ at the strong edge. Even though μ improved, the average diffraction efficiency η_{av} decreased according to the theory of section 4.2, since M increased, and therefore other noise sources degrade the total SNR. If, however, we were to keep $t_0/\tau_e = 0.011$ for the Fresnel filter, then η_{av} would improve, but μ would drop to 66.61 and 78.8 at the weak and strong edges, respectively.

4.4 Surface storage density of shift-multiplexed holographic 3–D disks

The surface storage density of a holographic disk is defined [44] as the number of bits of information (in the form of binary pixels) that are stored per unit area. Data is

stored so that every page, containing $N_p \times N_p$ pixels (N_p per dimension), occupies area $\mathcal{A}$ on the disk. In volume holographic memories, the page density is multiplied by M, the number of overlapping holograms per location. Therefore, the surface storage density $\mathcal{D}$ of any holographic disk is

$$\mathcal{D} = \frac{MN_p^2}{\mathcal{A}}. \tag{4.28}$$

The storage density for angle– and wavelength–multiplexed disks was calculated and optimized in [44]. In this section we will do the analogous calculation for shift–multiplexed disks when a spherical wave is used as reference.

First we consider the case when holograms are stored in the image plane. Specifically, we assume that the central pixel of the stored page is imaged at the center of the holographic medium. We denote by b the size of the pixels in the image. Then the area is $\mathcal{A} = (N_p b)^2/\cos\theta_S$ where θ_S is the angle of incidence of the central signal component, as in the previous sections. The number of overlapping shift multiplexed holograms along a single page is $M = N_p b/2\delta\cos\theta_S$, where δ is the shift selectivity given by (3.67), and we assume that successive holograms are stored at the 2nd Bragg null. This was justified in section 4.1. Therefore, we obtain for the density:

$$\mathcal{D}_{\text{image}} = \frac{N_p}{2b\delta} = \frac{N_p}{2b\lambda\left(\dfrac{z_0}{L\tan\theta_S} + \dfrac{1}{2(\text{NA})}\right)}. \tag{4.29}$$

For Fourier plane storage, the size of the first lobe (which contains all the information, according to the sampling theorem) is $2\lambda F/b$, where F is the focal length of the Fourier-transforming lens and b is the pixel size. The lobe size was derived assuming intensity detection. The result for the density is:

$$\mathcal{D}_{\text{Fourier}} = \frac{N_p^2 b}{4\lambda F\delta} = \frac{N_p^2 b}{4\lambda^2 F\left(\dfrac{z_0}{L\tan\theta_S} + \dfrac{1}{2(\text{NA})}\right)}. \tag{4.30}$$

Equations (4.29) and (4.30) give the density provided the distance z_0 has been

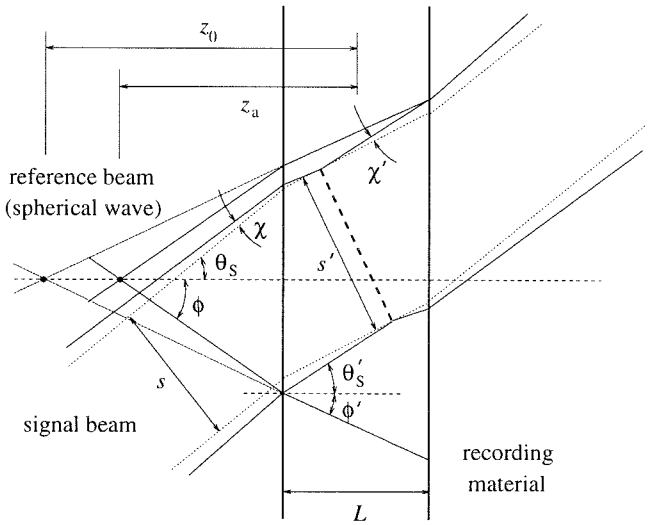


Figure 4.10: Geometry for the calculation of storage density in shift multiplexing geometry (spherical reference incident normally on the material, signal incident off–axis). The case $s' \sin\theta'_S < L$, $\phi < \theta_S$ is shown (see text).

already selected properly such that the reference and signal completely overlap inside the volume of the recording material. In general, the minimum z_0 is determined in terms of the hologram thickness and the geometry. We will show that z_0 varies linearly with L, according to the relation

$$z_0(L) = A + BL. \tag{4.31}$$

Increasing the thickness beyond a certain point does not lead to the expected gain in density, because the reduction in Bragg selectivity due to the increased interaction length competes with the simultaneous increase in z_0. In what follows, we will derive the coefficients A, B of (4.31) and the maximum achievable density with optimally selected z_0, as functions of thickness L.

We use ϕ for the angular spread of the reference beam, i.e., NA $= \sin\phi$. We will do the calculation simultaneously for the image and Fourier planes. For this reason,

we use the symbol s for the page size in both cases, given respectively by

$$s = \begin{cases} N_p b & \text{image plane} \\ \\ 2\lambda_0 F/b & \text{Fourier plane} \end{cases}, \tag{4.32}$$

where λ_0 is the wavelength of light in vacuum. The angular spread of the signal beam outside the holographic material is

$$\sin\chi \approx \begin{cases} \lambda_0/b & \text{image plane} \\ \\ N_p b/2F & \text{Fourier plane} \end{cases}. \tag{4.33}$$

Let n_0 denote the refractive index of the holographic material. The reference spread ϕ', the angle of signal incidence θ'_S, the signal spread χ' and the page size s' inside the material are recalculated using Snell's law as follows:

$$\sin\phi = n_0 \sin\phi', \tag{4.34}$$

$$\sin\theta_S = n_0 \sin\theta'_S, \tag{4.35}$$

$$\sin\chi = n_0 \sin\chi', \tag{4.36}$$

$$s\cos\theta'_S = s'\cos\theta_S. \tag{4.37}$$

Because the signal beam is tilted with respect to the normal to the recording material, it is possible that the tilted image of the data page does not fit inside the medium. This will happen if the medium is very thin and/or the tilt is large enough. Therefore, we need to consider two separate cases for thick and thin media. We start with the case of a thick medium so that the condition $s'\sin\theta'_S < L$ is satisfied, i.e., the whole focused page fits inside the hologram, as shown in Fig. 4.10. The geometry we chose for this analysis is conservative in the sense that we restricted the reference aperture according to $\phi < \theta_S$. This guarantees that the signal eventually separates itself from the reference cone, and thus the design of the imaging system that delivers

the signal to the hologram is simplified. This restriction could be relaxed, and the density would increase, but the optical design would become more complicated. We will not consider this optimization problem in this thesis.

A geometrical calculation based on Fig. 4.10 shows that the minimum z_0 required for the reference and signal to overlap is given by (4.31) with coefficients

$$A = \frac{s}{2}\frac{\cos\chi'\cos\theta'_S}{\cos(\theta'_S+\chi')\tan\phi'\cos\theta_S} \tag{4.38}$$

$$B = \frac{1}{2}\frac{\tan(\theta'_S+\chi')}{\tan\phi'}. \tag{4.39}$$

Recall that z_0 is the apparent focal distance of the spherical wave, as seen by an observer inside the holographic medium. In order to convert z_0 to z_a (focal distance measured in air), we apply (3.68). In the case of a thick medium, z_0 always increases with L, and the surface density saturates to

$$\mathcal{D}^{\max}_{\text{image}} = \frac{n_0 N_p \sin\phi'}{\lambda_0 b\left(1+\dfrac{\tan(\theta'_S+\chi')\cos\phi'}{\tan\theta'_S}\right)} \tag{4.40}$$

for the image plane, and

$$\mathcal{D}^{\max}_{\text{Fourier}} = \frac{n_0 N_p^2 b \sin\phi'}{2\lambda_0^2 F\left(1+\dfrac{\tan(\theta'_S+\chi')\cos\phi'}{\tan\theta'_S}\right)} \tag{4.41}$$

for the Fourier plane.

For the case $s'\sin\theta'_S > L$ (thin medium), the geometrical calculation is more complicated. Using the same restrictions as before, the result is:

$$A = \frac{s}{2}\frac{\cos\chi}{\tan\phi'\cos(\theta_S+\chi)} \tag{4.42}$$

$$B = \frac{1}{2\tan\phi'}\left[\frac{1}{\tan\theta'_S}+2\tan(\theta'_S+\chi')-\frac{\cos\theta_S\cos\chi}{\sin\theta'_S\cos\theta'_S\cos(\theta_S+\chi)}\right]. \tag{4.43}$$

Note that the coefficient B for a thin medium may become negative, and then the

optimal z_0 decreases with L. This is due to the bending of the signal rays induced by refraction.

In a recent experiment [33], surface storage density in excess of 10bits/μm^2 with raw bit error rate of 10^{-4} was demonstrated in a holographic disk configuration using Du Pont's HRF–150–100 photopolymer as the recording material. The parameters used in this experiment were $\lambda = 532$ nm, $n_0 = 1.525$, $N_p = 768$, $b = 45\mu$m, $F = 5.46$ cm. 32 Fresnel region holograms were superimposed on the same spot using a combination of angle (8 locations, separated by 4 Bragg nulls) and peristrophic [41] (4 holograms per angular location) multiplexing.

Shift multiplexing can also be combined with other techniques, such as peristrophic and fractal [30], in order to increase the storage density at the cost of complicating page access. Better yet, it is possible to apply the spherical–reference analog to the angle plus fractal/peristrophic methods, which consists of shift multiplexing holograms in both x and y directions (see Fig. 3.11). In the disk configuration, y–shift multiplexing corresponds to overlapping hologram tracks. The y–shift selectivity for high–bandwidth signal beams was calculated in section 3.3.3. For the same parameters of the experiment of [33], y–shift multiplexing increases the density by a factor of at least 3.

Using the combination of y–shift multiplexing with x–shift multiplexing at the 4th shift Bragg null (consistent with [33]), $\phi = 45°$, $\theta_S = 60°$, and assuming Fourier plane storage, we obtain the theoretical density prediction for shift multiplexing, given in Fig. 4.11. Notice that, for the thickness $L = 100\mu$m of the Du Pont photopolymer, $x + y$–shift multiplexing is expected to yield $\mathcal{D} = 11.8$bits/μm^2, slightly higher than the 10.7bits/μm^2 of the high density experiment reported in [33]. This result has since been verified experimentally [93, 94]. The point ($L = 1$ mm, $\mathcal{D} = 100$bits/μm^2) has also been verified experimentally [87]. Shift density increases almost linearly with thickness, reaching 163.4bits/μm^2 for $L = 1.2$ mm, when it begins to saturate. Thus shift multiplexing utilizes the area of holographic 3D disks more efficiently.

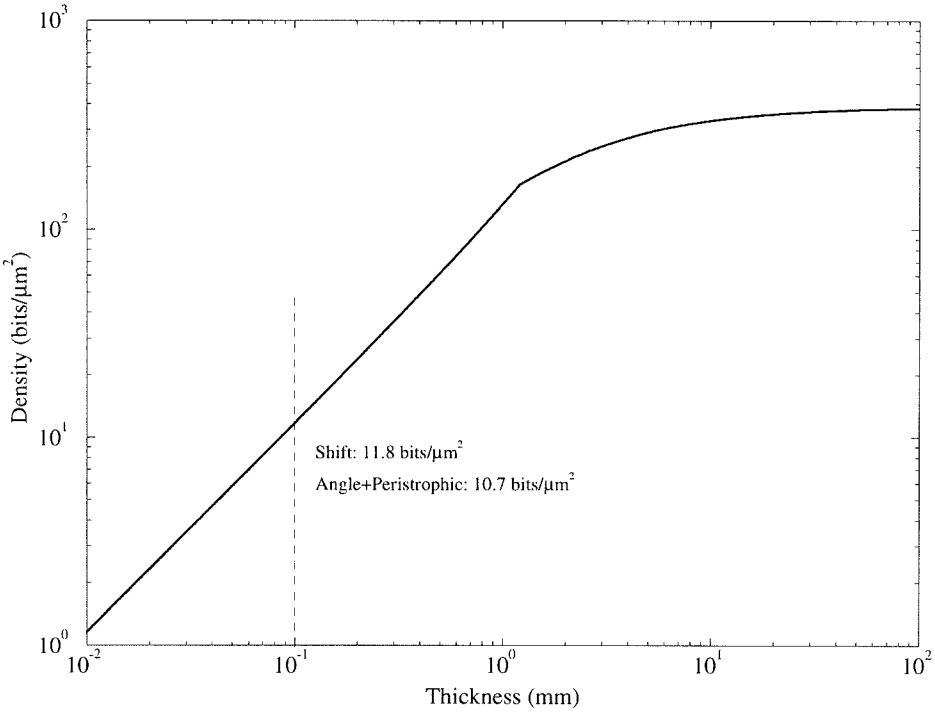


Figure 4.11: Theoretical shift multiplexing surface storage density in the Fourier plane, using parameters $\lambda = 0.532$ nm, $n_0 = 1.525$, $N_p = 768$, $b = 45\mu$m, $F = 5.46$ cm, consistent with the angle+peristrophic experiment. The reference spread used for the shift multiplexing density calculation is $\phi = 45°$, and the signal beam is incident at $\theta_S = 60°$.

4.5 Readout with slow erasure

When photorefractive crystals are used to implement rewritable shift multiplexed memories, an issue of concern is the undesired erasure of the recorded holograms during read–out. Techniques have been developed for fixing holograms thermally [95] or electrically [96, 97], and sustaining holograms by periodic refreshing [98] (see section 6.1). Here we consider the two–lambda method, in which the recording and read–out wavelengths are different [99]. The allowable read–out time is prolonged if the read–out wavelength is selected such that the absorption is lower and hence erasure is reduced. Because the Bragg matching condition at the new wavelength is modified, only a portion of the angular spectrum of the recorded holograms can be Bragg–matched with a single plane wave. It is possible, however, to reconstruct in its entirety a single hologram recorded with a plane wave, by reading it out with a spherical beam [82]. In this section, we show how the two–lambda technique ap-

plies to shift multiplexed memories, where holograms are recorded and read–out with spherical wave references.

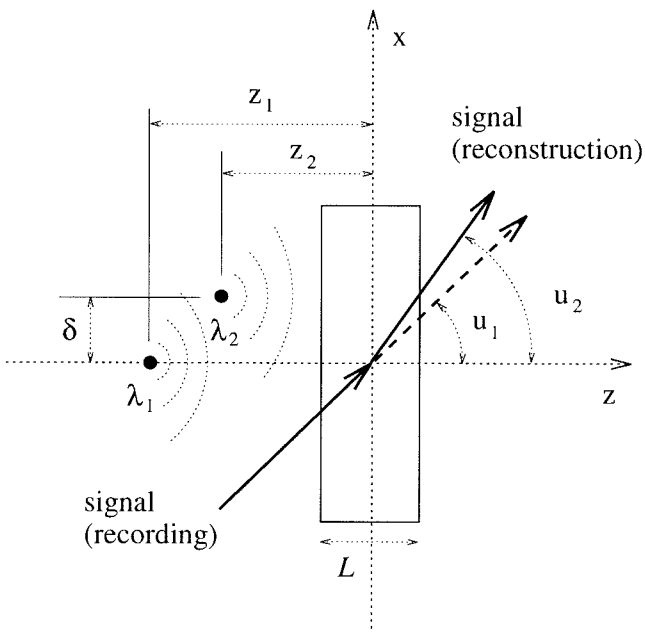


Figure 4.12: Geometry for the two–lambda technique with shift–multiplexing.

We will use subscript 1 for quantities associated with the recording wavelength λ_1 and subscript 2 for the read–out wavelength λ_2. The geometry is shown in Fig. 4.12. The holographic material occupies the region $-\infty < x, y < +\infty$, $|z| < L/2$. The recording reference beam (at wavelength λ_1) is a spherical wave originating from $(0, 0, -z_1)$. For now, we consider the signal beam to be a single plane wave component propagating at angle $u_1 \approx \sin u_1$ (paraxial approximation) with respect to the hologram normal. The read–out spherical wave (at wavelength λ_2) originates at $(\delta, 0, -z_2)$. Under the Born, paraxial, and constant modulation depth approximations [70, 43] the diffracted field can be calculated analytically (see also the Appendix to section 3.3.1). The reconstruction is a plane wave if

$$\lambda_1 z_1 = \lambda_2 z_2. \tag{4.44}$$

This change in focal distance compensates for the curvature mismatch between the

recording and read–out reference wavefronts. The plane wave hologram is Bragg matched at the read–out wavelength, if it is translated relative to the read–out reference by

$$\delta_{\mathrm{B}} = \frac{1}{2}\left(\frac{\lambda_2}{\lambda_1} - 1\right) u_1 z_2. \tag{4.45}$$

If Eqs. (4.44) and (4.45) are both satisfied, the reconstruction is a plane wave propagating at angle u_2 satisfying (in the paraxial approximation) the condition

$$\frac{u_2}{u_1} = \frac{\lambda_2}{\lambda_1}. \tag{4.46}$$

Once the hologram is Bragg matched after shifting by δ_{B}, an additional translation by the shift Bragg selectivity

$$\delta_2 = \frac{\lambda_2 z_2}{u_2 L} = \frac{\lambda_1}{\lambda_2}\delta_1 \tag{4.47}$$

will eliminate the reconstruction. A separate plane wave hologram recorded at that shifted position can be observed without interference. In Eq. (4.47), $\delta_1 = \lambda_1 z_1/(u_1 L)$ is the shift selectivity at the recording wavelength [43]. Note that if $\lambda_2 > \lambda_1$, then δ_2 is smaller than δ_1. Since holograms are recorded δ_2 apart, packing is more dense compared to the case when read–out is intended to be at wavelength λ_1.

In the analysis so far we have neglected refraction and dispersion effects, and the finite numerical apertures $(\mathrm{NA})_1$, $(\mathrm{NA})_2$ of the spherical waves. The properly modified formulas are given in Table 4.1, where the primed quantities are measured in air, and n_1, n_2 are the refractive indices. In general, $(\mathrm{NA})_2 < (\mathrm{NA})_1$ because the required shift δ_{B} introduces edge effects.

These theoretical predictions were verified experimentally using a thin Fe–doped $LiNbO_3$ crystal grown by Deltronics. The experimental parameters were: $\lambda_1' = 488$ nm, $\lambda_2' = 632.8$ nm, $u_1' \approx 40° = 0.698$ rad, $n_1 = 2.3533$, $n_2 = 2.2935$, $L = 250$ μm, $(\mathrm{NA})_1 \approx 0.1$. The quoted values of refractive index were calculated by interpolating the data given in [100]. This experiment was done with a plane wave signal beam.

Bragg selectivity	δ_1	$\frac{\lambda_1'[n_1 z_1' - (n_1-1)L/2]}{u_1' L} + \frac{\lambda_1'}{2(\mathrm{NA})_1}$
	δ_2	$\frac{\lambda_2'[n_2 z_2' - (n_2-1)L/2]}{u_2' L} + \frac{\lambda_2'}{2(\mathrm{NA})_2}$
Focusing condition		$\lambda_1'\left(z_1' - \left(1-\frac{1}{n_1}\right)\frac{L}{2}\right) = \lambda_2'\left(z_2' - \left(1-\frac{1}{n_2}\right)\frac{L}{2}\right)$
Bragg–matching condition	δ_B	$\frac{1}{2}\left(\frac{n_1\lambda_2'}{n_2\lambda_1'} - 1\right)\frac{u_1'}{n_1}\times \left[n_2 z_2' - (n_2-1)\frac{L}{2}\right] + \frac{\lambda_2'}{2(\mathrm{NA})_2}$
Reconstructed band	Δu_2	$\frac{\lambda_1'}{\left(\frac{n_1\lambda_2'}{n_2\lambda_1'} - 1\right) u_{1\mathrm{c}}' L}$

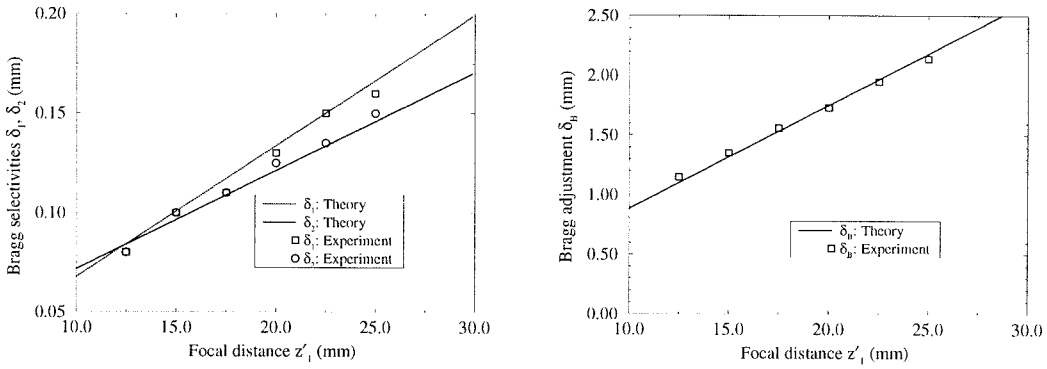
Table 4.1: Two–lambda equations including refraction and dispersion.

The measured values of δ_B, δ_1, δ_2 are plotted in Fig. 4.13 together with the theoretical predictions derived from the formulas of Table 4.1. $(\mathrm{NA})_2$ was estimated to 0.014. The measurements for λ_2 read–out were made by first refocusing the read–out beam according to the theoretical prediction for z_2'.

An information–bearing signal occupies a finite angular bandwidth around the carrier $u_{1\mathrm{c}}$. Since δ_B (eq. 4.45) is different for each spatial frequency component, the entire hologram cannot be Bragg–matched at the same time. The reconstruction consists of the portion of the angular spectrum lying in the range

$$\Delta u_2 = \pm \frac{2\lambda_1}{\left(\frac{\lambda_2}{\lambda_1} - 1\right) u_{1\mathrm{c}} L}. \tag{4.48}$$

For Fourier transform holograms, Eq. (4.48) implies that a slice of size $\Delta x = 2\Delta u_2 F$



(a) (b)

Figure 4.13: Experimental results for the Bragg matching and selectivity properties of the two–lambda method applied to shift multiplexed holograms.

(where F is the focal length of the lens used for the Fourier transformation) is reconstructed. For Image plane holograms, the reconstruction is low–pass filtered with a cut-off $\Delta w = \Delta u_2/\lambda_2$. The modified version for Eq. (4.48) accounting for dispersion is given in Table 4.1.

We used the same experimental setup to record a single Fourier plane hologram (focal length $F = 20$ cm) of a transparency with z'_1=22.5 mm. The reconstruction at wavelength λ_1 is shown in Fig. 4.14(a). Two reconstructions at λ_2 (obtained by changing δ_B) are shown in Fig. 4.14(b,c). As predicted, only a slice of the stored image is obtained at one time. The width of the slice was measured to be 0.7 mm, in agreement with Eq. (4.48). If we continuously shift the medium, a sliding window of the stored image will appear on the CCD. In addition, the reconstruction as viewed through this sliding window is also shifting due to the motion of the medium. A time-delay-and-integrate (TDI) detector array can compensate for this motion and integrate the response to produce a complete, unblurred image. A TDI detector was simulated in software to produce the complete reconstruction shown in Fig. 4.14(d). Alternatively, we can set the size of the recorded images in the x-direction equal to the width of the reconstructed slice. Either method results in the complete reconstruction of the stored images at the expense of reduced storage density.

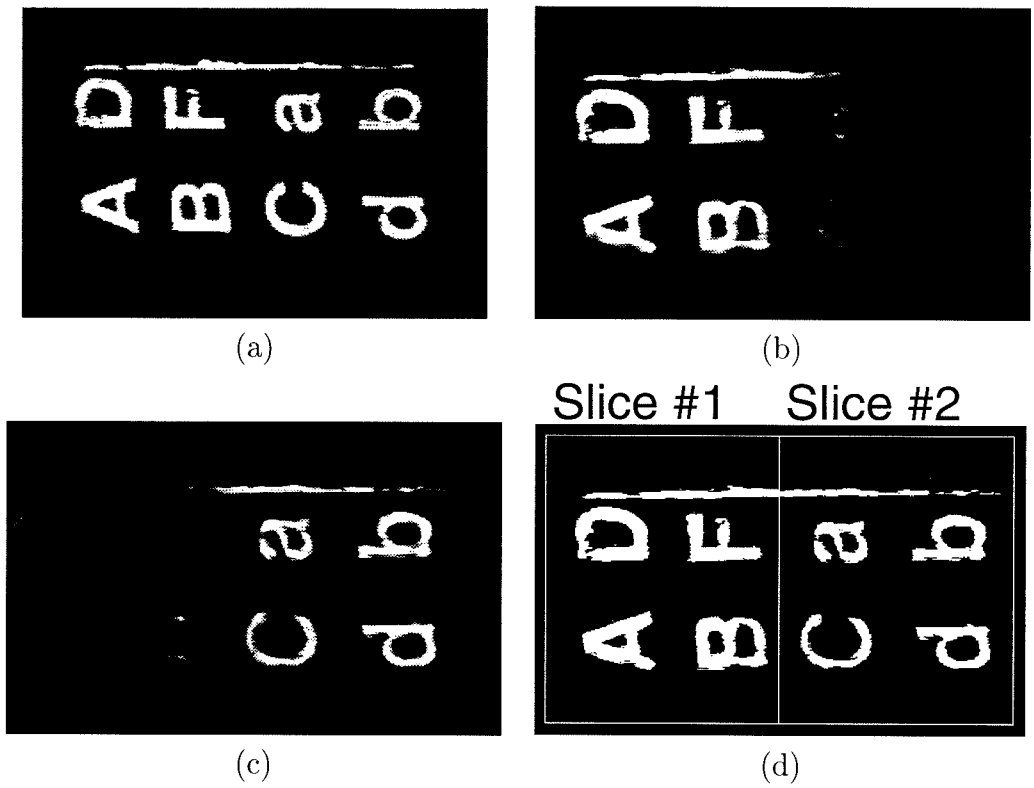


Figure 4.14: Hologram reconstructions obtained with the two–lambda method.

The surface storage density of a shift multiplexed memory can be expressed as:

$$\mathcal{D} = \frac{N_{px} N_{py}}{s_1 \delta_2}, \tag{4.49}$$

where N_{px} and N_{py} are the number of SLM pixels in the x and y directions, respectively, and s_1 is the transverse size of the signal beam. For Fourier plane recording, we can relate N_{px} to the geometry of the holographic system by imposing the requirement that the angular bandwidth of the stored holograms (and hence the hologram thickness) is matched to the width of the allowable reconstructed angles (the SLM size). Substituting Eq. (4.48) and Eq. (4.49), we obtain:

$$\mathcal{D} = \frac{2N_{py}}{\left(\frac{\lambda_2}{\lambda_1} - 1\right) \lambda_1 z_1}. \tag{4.50}$$

The above equation is valid only if the SLM size is large enough to accommodate the

entire strip. Notice that when $\lambda_2 = \lambda_1$, the strip is infinite, and therefore Eq. (4.50) does not hold. As an example, if $N_{py} = 1,000$, $z_1' = 1.3$ mm, $\lambda_1' = 488$ nm, $\lambda_2' = 633$ nm, then $\mathcal{D} = 10.6\text{bits}/\mu\text{m}^2$. A similar derivation can be carried out for the density of the Image plane geometry and it leads to the same expression as Eq. (4.50).

Chapter 5 Imaging systems for holographic memories

5.1 The two basic imaging systems

5.1.1 The 4–F imaging system

The 4–F system is a simple, commonly used optical system for imaging. It is comprised of two lenses with focal length F, separated by distance $2F$. The input is placed F in front of the first lens and illuminated with a plane wave. Then the Fourier transform of the input appears mid-way between the two lenses. The second lens applies yet one more Fourier transform, and the input, reversed, is obtained. In the context of holographic memories, imaging is needed to re-focus the reconstruction on the detector (e.g., a CCD camera). The hologram can be in principle placed anywhere between the spatial light modulator (SLM) and CCD planes; however, it makes sense to place it at the focal points to minimize the area occupied in the storage medium. Then, depending on the exact location of the hologram, we obtain two different geometries, Fourier and Image plane, shown in Figure 5.1.

In the case of Fourier plane storage, there is a caveat to placing the hologram exactly at the Fourier plane: because the signal comes to a sharp focus, the intensity non-uniformity during recording degrades severely the quality of the holograms. Thus it is necessary to defocus the hologram, at the expense of areal storage density. The remaining analysis that will follow is still valid for the defocused holograms.

To understand the trade-offs involved in using the 4–F system for holographic storage, we must study the effects of imaging restrictions (i.e., the requirement of resolvable reconstruction at the output plane) on the storage density and crosstalk. Here we state the imaging restrictions for the case of 4–F systems. Consider the

subset of the imaging system shown in Figure 5.2.

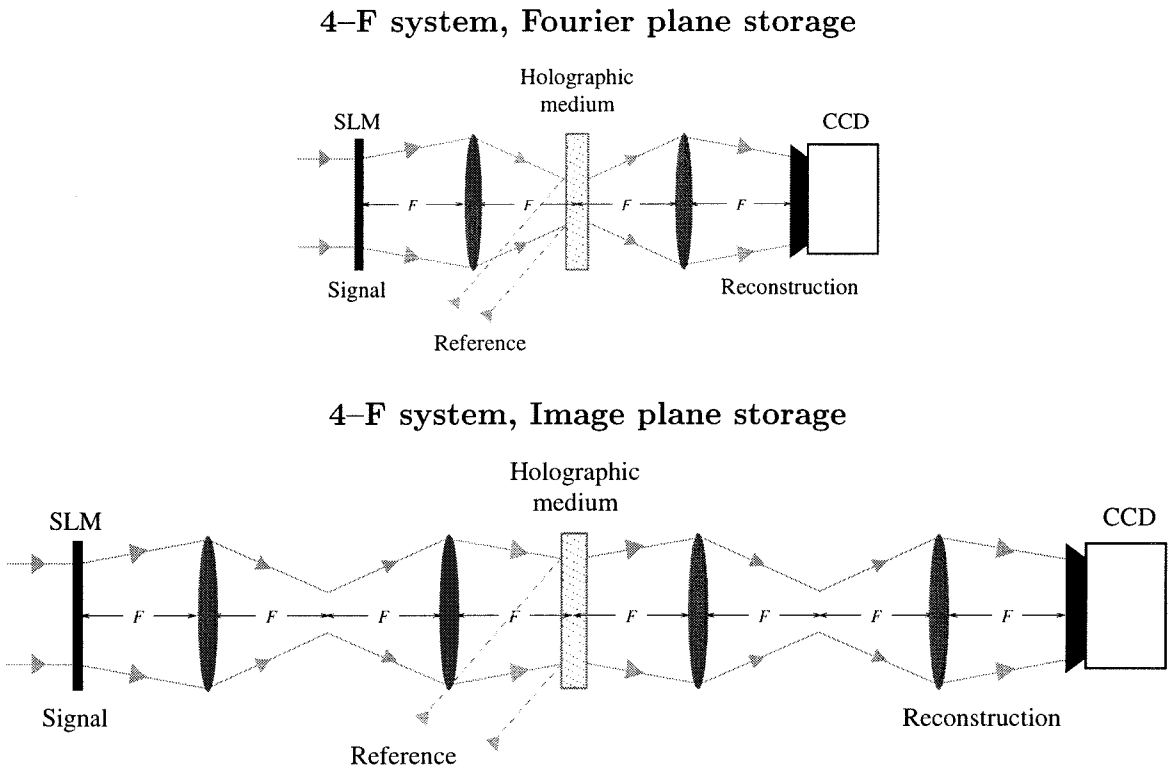


Figure 5.1: Holographic storage architectures with the 4–F imaging system.

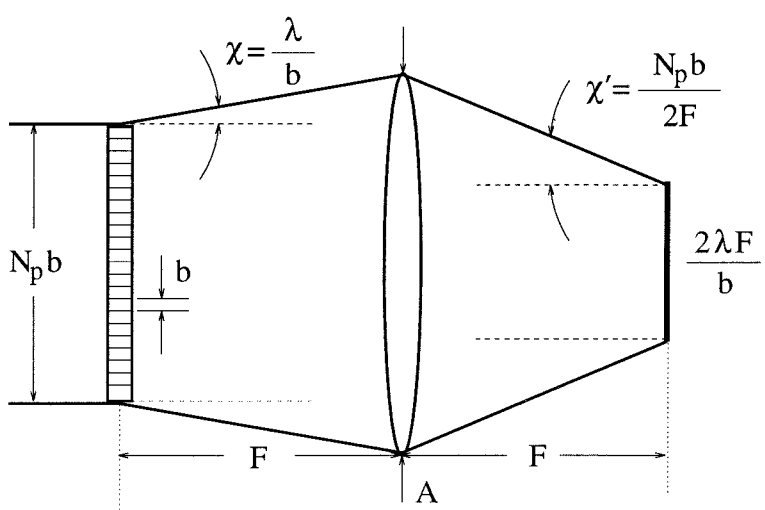


Figure 5.2: Imaging restrictions on a 4–F system.

The SLM has N_p pixels, and each has size b (for simplicity, we assume that the fill-factor is 1). The lens has focal length F, and aperture A (therefore, the F-number is $(F/\#) = F/2A$). The diffraction spreads before and after the lens are, respectively

$$\chi = \frac{\lambda}{b} \qquad \chi' = \frac{N_p b}{2F} \tag{5.1}$$

Note that at the second arm of the 4–F system, χ' and χ are interchanged. The imaging restriction follows from the requirement that the entire diffraction lobe contained in χ, χ' be contained in the lens aperture. Thus we obtain

$$N_p b + \frac{2\lambda F}{b} \leq A, \tag{5.2}$$

with equality in the optimal case of full utilization of the system aperture (beyond some point, this degrades image quality because of aberrations at the lens edges; for a detailed account of these issues, see [101, 102]). From (5.2) we obtain the following conditions:

$$N_p \leq N_{p,\max} \equiv \frac{A^2}{8\lambda F} \tag{5.3}$$

$$b_{\min} \leq b \leq b_{\max} \tag{5.4}$$

$$b_{\min} = \frac{4\lambda F}{A}\frac{N_{p,\max}}{N_p}\left(1 - \sqrt{1 - \frac{N_p}{N_{p,\max}}}\right) \tag{5.5}$$

$$b_{\max} = \frac{4\lambda F}{A}\frac{N_{p,\max}}{N_p}\left(1 + \sqrt{1 - \frac{N_p}{N_{p,\max}}}\right) \tag{5.6}$$

According to these conditions, the parameters N_p, b are not free, but are constrained to belong to a parabolic-shaped region, as shown in Figure 5.3.

5.1.2 The van der Lugt imaging system

Observing the schematics for the 4–F imaging system, it is natural to wonder if it might be possible to reduce the diffraction spread by illuminating the SLM with a focused wavefront rather than a plane wave. This method would attain more efficient

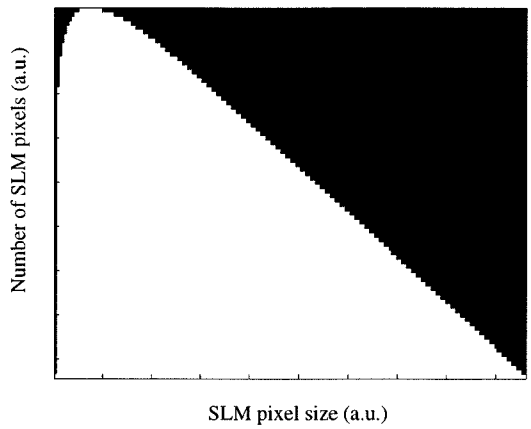


Figure 5.3: Shape of the allowable region for N_p, b in a 4–F system.

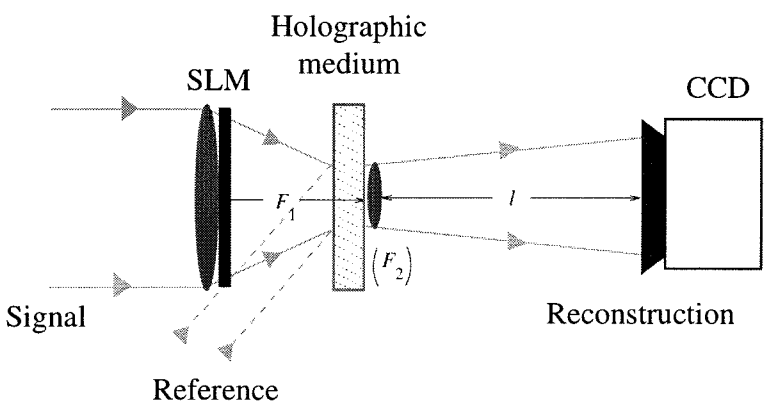


Figure 5.4: Imaging system proposed by van der Lugt.

usage of the system aperture. The answer was given by van der Lugt [103], who proposed the architecture shown in Figure 5.4.

The imaging condition and magnification of this system are calculated using simple Fourier optics, and are given by:

$$\frac{1}{F_1} + \frac{1}{l} = \frac{1}{F_2} \qquad |M| = \frac{F_1}{F_2} - 1 = \frac{F_1}{l} \tag{5.7}$$

Like in the case of the 4–F system, we must require that the first diffraction lobe from the SLM pass completely through the apertures. This yields the following

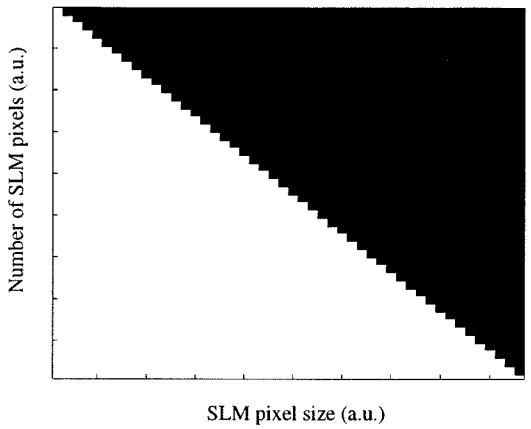


Figure 5.5: Shape of the allowable region for N_p, b in a van der Lugt system.

constraints:

$$\frac{A_2}{F_2} = (1 + |M|)\frac{2\lambda}{b} \tag{5.8}$$

$$N_p \leq \frac{F_1(F/\#)}{b} \leq \frac{F_1(F/\#)^2}{2\lambda(1 + |M|)} \tag{5.9}$$

$$b_{\min} \leq b \leq b_{\max} \tag{5.10}$$

$$b_{\min} = \frac{2\lambda(1 + |M|)}{(F/\#)} \tag{5.11}$$

$$b_{\max} = 2F_1(F/\#) = A_1 \tag{5.12}$$

The resulting allowable region is larger than that of the 4–F system and has a triangular shape, as shown in Figure 5.5.

5.2 Angle-multiplexed memories

We will now compare the three imaging architectures (Fourier plane, image plane, and van der Lugt) described in the previous section for the case of an angle-multiplexed memory. The comparison is usually made on the basis of two metrics: surface storage density, and crosstalk. We will see that optimizing one of the two contradicts the optimization of the other. Then using the information capacity metric developed in

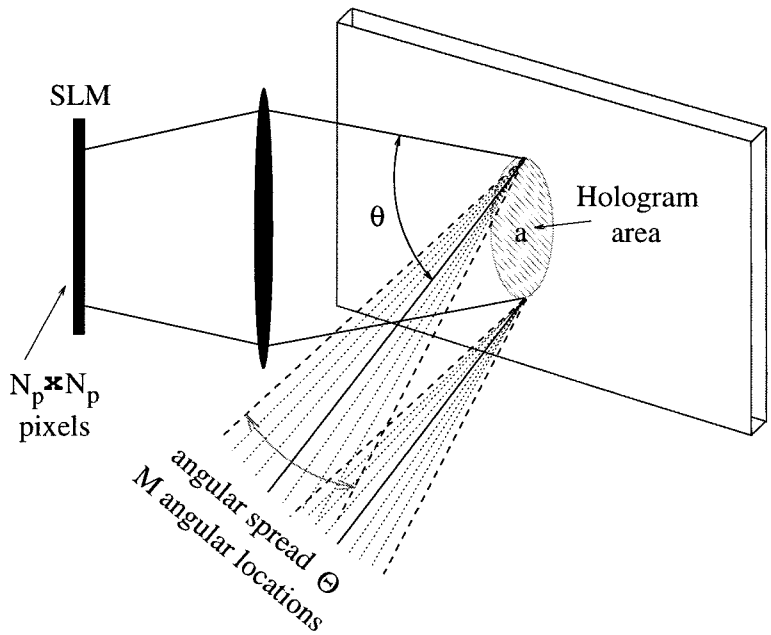


Figure 5.6: Angle-multiplexed system used for the subsequent storage density and crosstalk calculations.

section 2.3, we will unify the two metrics, making the comparison easier.

The architecture is described schematically in Figure 5.6. The signal is normally incident on the holographic material, which has thickness L. The reference beam is incident at angle θ, and M angular locations are available for multiplexing. If $\Delta\theta$ is the angular separation between adjacent holograms (e.g., equal to one, two or more Bragg nulls) and Θ is the total angular spread of the reference, then

$$M = \frac{\Theta}{\Delta\theta}. \tag{5.13}$$

Here and in the subsequent analysis we will ignore for simplicity refraction effects and lens aberrations.

5.2.1 Raw surface storage density

We assume that adjacent, spatially multiplexed holograms are separated by the area a occupied by the signal beam at the surface of the holographic material, as shown

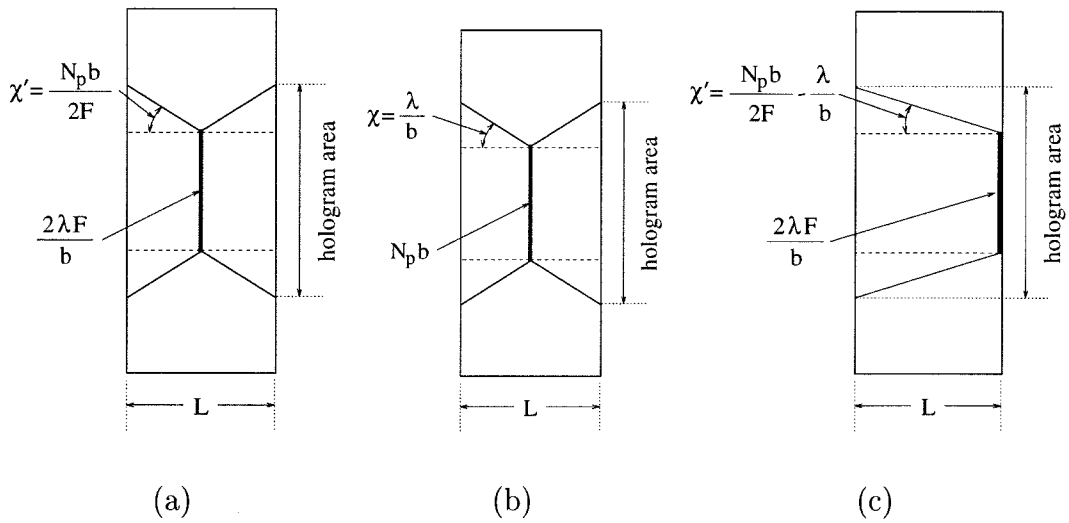


Figure 5.7: Geometry for the calculation of the hologram area in a 4–F holographic storage system: (a) Fourier plane geometry, (b) image plane geometry, and (c) the van der Lugt system.

in Figure 5.7. This ignores the area taken by the reference (which is incident at an angle; see Fig. 5.6). In principle, this is not a problem. However, if the reference illuminates adjacent holograms at any stage during recording or reconstruction, it will erase them, thereby reducing the overall dynamic range of the system. We will not consider this effect here.

The raw[1] surface storage density is defined in terms of the area a occupied by the hologram (see Figure 5.6) as

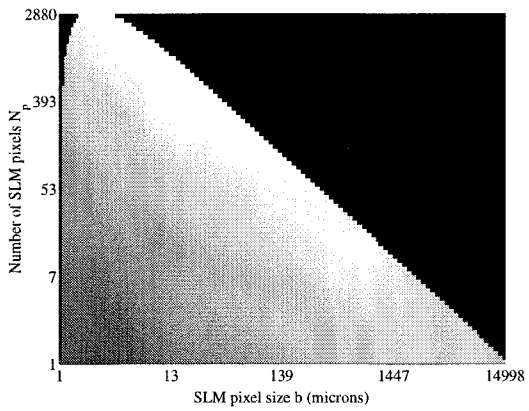
$$\mathcal{D} = \frac{MN_p^2}{a}. \tag{5.14}$$

The geometrical calculation (Fig. 5.7) of the area yields the following results:

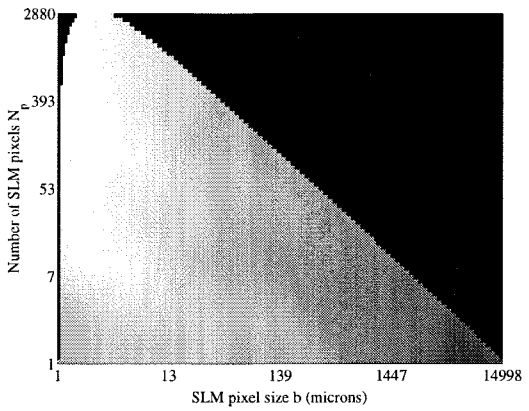
$$a_{\mathrm{F}} = \left(\frac{2\lambda F}{b} + \frac{N_p b L}{2F}\right)^2, \qquad a_{\mathrm{I}} = \left(N_p b + \frac{\lambda L}{b}\right)^2, \qquad \text{and}$$

$$a_{\mathrm{V}} = \left[\frac{2\lambda F}{b} + 2\left(\frac{N_p b}{2F} - \frac{\lambda}{b}\right) L\right]^2,$$

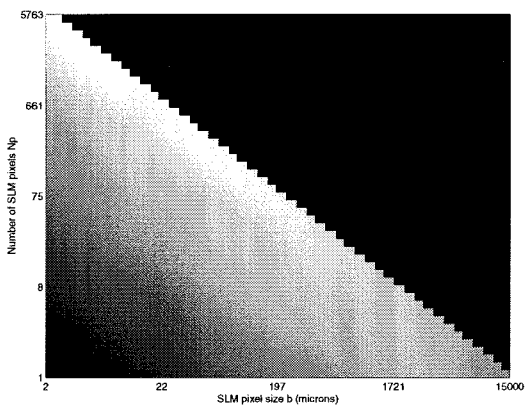
[1] *Raw* density refers to the number of bits that can be stored before error correction. An upper bound for the capacity, taking noise into account, will be derived in section 5.2.3.



(a)



(b)



(c)

a) 4–F system, Fourier plane.

b) 4–F system, Image plane.

c) van der Lugt system.

Figure 5.8: Surface storage density versus N_p and b. The brightness in the images is proportional to the density. The calculation was made for material thickness $L = 100\mu$m, angle of incidence $\theta = 30°$, reference angular spread $\Theta = 20°$, focal length $F = 20$ cm, and lens aperture $A = 15$ cm.

where the subscripts "F," "I," "V" denote Fourier plane, Image plane and van der Lugt storage respectively. Substituting into the density equation (5.14), we can calculate the density for any pair N_p, b. An example is given in Figure 5.8 for the three cases of Image plane, Fourier plane and van der Lugt systems. We observe that (as perhaps was expected) the density in the 4–F Fourier plane and van der Lugt systems improves as the product $N_p b$ increases, while the 4–F image plane density is high for small b.

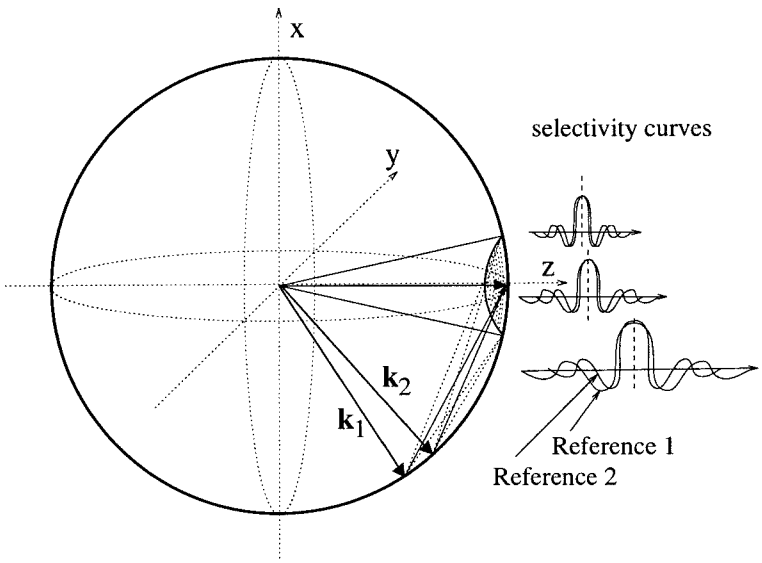


Figure 5.9: Explanation of crosstalk with the aid of the k-sphere (see also Figure 3.2): the reference beam with wavevector $\mathbf{k}_2$ always has Bragg selectivity curve narrower than $\mathbf{k}_1$, but the width depends on the position along the hologram.

5.2.2 Inter-page and intra-page crosstalk

Two sources of crosstalk have been identified in holographic memories: inter-page, and intra-page. Interpage crosstalk is explained in Figure 5.9: the Bragg selectivity is not the same along a wideband hologram. Therefore, when the carrier component of a hologram is Bragg-mismatched, some other components still diffract contributing to crosstalk. It follows that inter-page crosstalk affects worse the *edges* of a Fourier transform hologram, and the *high frequency components* of an image plane hologram.

The crosstalk for angle-multiplexed Fourier plane holograms was derived in [71]. The electric field amplitude of the reconstruction of the j-th hologram is

$$g_j(x', y') = \sum_{m=0}^{M} f_m\left(-x' + \lambda Q_{mj}, -y'\right) \operatorname{sinc}\left((m-j)\left[C_{mj} + R_{mj}\frac{x'}{F}\right]\right), \tag{5.15}$$

where f_m are the stored images, (x', y'), the output coordinates, and the constants

are defined as

$$Q_{mj} = \frac{(m-j)F\cot\theta}{L} + \frac{(m^2-j^2)\lambda F}{L^2\sin\theta} \qquad R_{mj} = \cot\theta - \frac{(m+j)\lambda}{L\sin\theta}$$

$$C_{mj} = 1 - \frac{(m-j)\lambda\cos^2\theta}{L\sin\theta} + \frac{1}{2}\frac{(m+j)\lambda\cos\theta}{L\sin^2\theta}\left(1 - \frac{(m+j)\lambda}{L\sin\theta}\right).$$

The only desired term in the summation of (5.15) is the one with $m = j$. The remaining terms are crosstalk. By assuming that the crosstalk contributions from different holograms add *incoherently*, i.e., that the relative phases of the holograms are randomly distributed in $[0, 2\pi)$, we obtain the SNR as

$$\frac{1}{\mathrm{SNR}_{\mathrm{F,inter}}(x', y')} = \sum_{m\neq j} \mathrm{sinc}^2\left((m-j)\left[C_{mj} + R_{mj}\frac{x'}{F}\right]\right). \tag{5.16}$$

We observe that the argument to the sinc function of (5.16) contains two terms: one constant, due to the fact that the selectivities of the holograms are not all the same[2] (since they are recorded at different angles θ), and one that depends on the transverse location x', due to the fact that pixels located off-axis also have different selectivity than the carrier of the hologram (for a demonstration of this effect in the equivalent case of shift multiplexing, see Figures 4.2 and 4.3).

The expression derived above for Fourier-plane crosstalk holds approximately for the van der Lugt system, except the dependence on the transverse coordinate x'/F must be scaled by $(A_1/2F_1 - \lambda/b)$ to correct for the reduced signal beam spread at the Fourier plane. This in general results in lower crosstalk than the regular Fourier plane.

The crosstalk for low-bandwidth image plane holograms in the 90° geometry was calculated in [89]. Here we will modify the analysis for transmission geometry, and in order to take into account high-bandwidth images. The amplitude contribution is obtained simply by taking the Fourier transform of the Fourier plane result (5.15).

[2]Note that this problem does not exist in the case of shift and wavelength multiplexing: indeed, in the analysis of sections 4.1, 5.3, and 5.4, C_{mj} simplifies to 1.

Subsequently, we assume that the spatial cross-correlation of the stored holograms is

$$\mathcal{R}_{f_j f_{j'}}(\xi) = \delta_{jj'} \mathrm{sinc}\,(w\xi)\,, \tag{5.17}$$

where w is the bandwidth of the stored images (loosely defined as $1/b$). Then the spatial auto-correlation of the reconstruction of the j-th hologram is obtained as

$$\mathcal{R}_{g_j}(\xi) = \mathcal{R}_{f_j}(\xi) + \sum_{m \neq j} \int_{-1}^{+1} (1 - |\xi'|)\,\mathrm{sinc}^2\,(w\,[\lambda(m-j)R_{mj}\xi' - \xi]) \exp\{i2\pi(m-j)C_{mj}\xi'\}\,\mathrm{d}\xi', \tag{5.18}$$

where $\mathcal{R}_{f_j}(\xi)$ is the autocorrelation of the originally stored j-th hologram, given by (5.17). The variance added to the reconstruction due to crosstalk noise is obtained by calculating the autocorrelation at zero displacement, $\xi = 0$ (see also [104]). Therefore, we obtain:

$$\frac{1}{\mathrm{SNR}_{\mathrm{I,inter}}} = \sum_{m \neq j} \int_{-1}^{+1} (1 - |\xi'|)\,\mathrm{sinc}^2\,(\lambda w(m-j)R_{mj}\xi') \exp\{i2\pi(m-j)C_{mj}\xi'\}\,\mathrm{d}\xi'. \tag{5.19}$$

For small w (i.e., $b >> \lambda$), the above expression reduces to

$$\frac{1}{\mathrm{SNR}_{\mathrm{I,inter}}} \approx \sum_{m \neq j} \mathrm{sinc}^2\,((m-j)C_{mj})\,. \tag{5.20}$$

If we further assume $\theta = 90°$ (i.e., 90°-geometry), then we obtain the result derived by K. Curtis [89, 105]:

$$\frac{1}{\mathrm{SNR}_{\mathrm{I,inter}}} \approx \sum_{m \neq j} \mathrm{sinc}^2 \left((m-j) \left[1 - (m-j)(m+j)^2 \frac{\lambda^3}{L^3} \right] \right). \tag{5.21}$$

It is important to take the bandwidth into account, because typically (5.19) gives SNR 6-7 orders of magnitude lower than (5.20).

Intra-page crosstalk results from the finite point-spread function of the optical

system. A straightforward calculation [106] yields

$$\frac{2}{\mathrm{SNR}_{\mathrm{intra}}} \approx \sum_{n=-N_p/2}^{N_p/2} \left(\int \mathrm{rect}\left(\frac{x'' - nb}{b}\right) \mathrm{sinc}\left(2\frac{x + x''}{b}\right) \mathrm{d}x'' \right)^2 . \tag{5.22}$$

Assuming Gaussian noise distributions, the overall SNR is

$$\frac{1}{\mathrm{SNR}_a} = \frac{1}{\mathrm{SNR}_{a,\mathrm{inter}}} + \frac{1}{\mathrm{SNR}_{\mathrm{intra}}} \tag{5.23}$$

for a=F, V, or I. Using all the results so far, we can perform the SNR calculation for every pair of N_p, b as in the case of surface density. An example is given in Figure 5.10.

We notice immediately that the high SNR areas in all architectures correspond to low surface storage densities, and vice versa. Thus to pursue high raw density, one should expect to pay the price of low SNR. The overhead of error correction in the case of strong noise reduces the effective density. The trade-off is quantified with the aid of the concept of information density (defined in section 2.3) in the next section.

5.2.3 Information density

The information density $\mathcal{C}$ of a holographic memory is determined with the help of the noise theory presented in Chapter 2. The density calculation yields the number of raw bits that can be stored per unit area in the memory. The crosstalk calculation yields the SNR, from which the error rate is deduced according to section 2.2. Finally, the information metric of section 2.3 provides the trade-off between the raw density and noise performance as the upper limit in useful information that can be stored per unit area in the memory, according to Shannon's theorem. The result for the case considered in the previous two sections (Figures 5.8 and 5.10) is given in Figure 5.11. We observe that the plots are very similar to the density plots, but the information density is always smaller than the raw density. Additional noise sources that do not depend on the parameters b, N_p of the imaging system and were neglected would

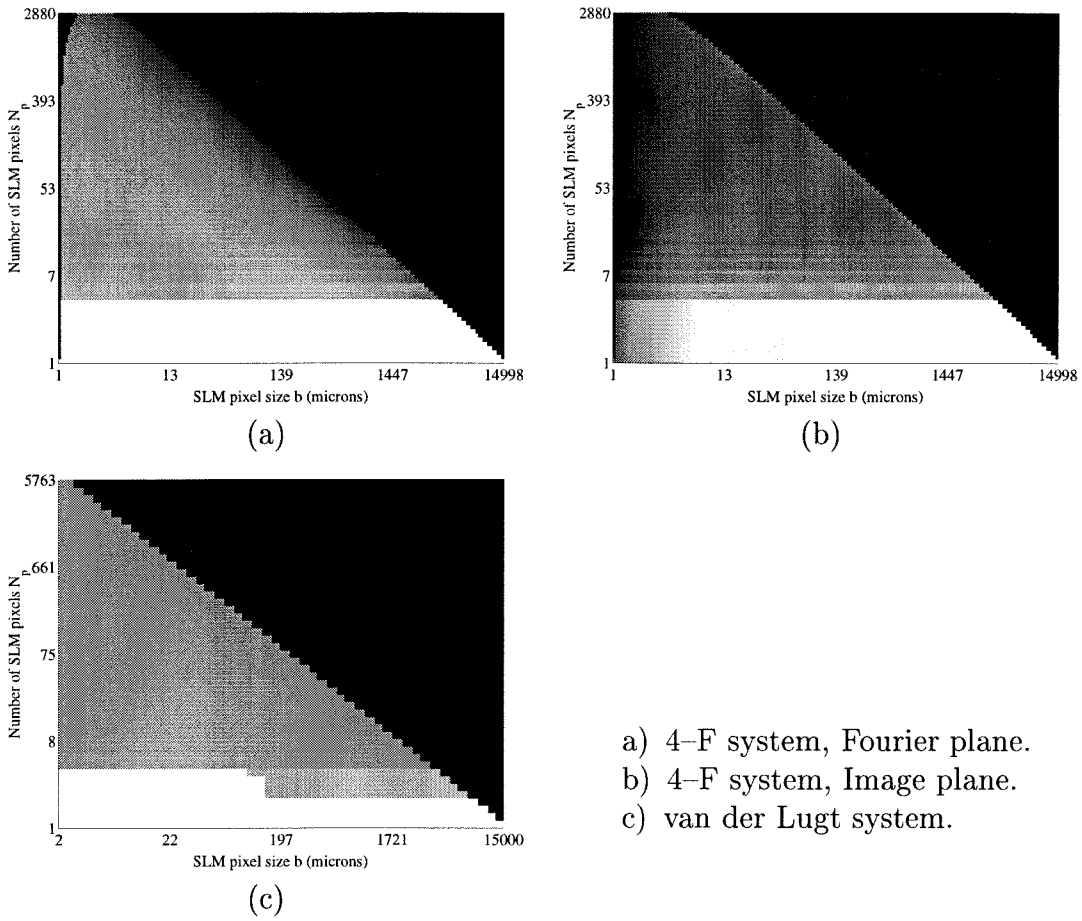


Figure 5.10: Combined inter- and intra-page crosstalk SNR versus N_p and b. The brightness in the images is proportional to the SNR. The calculation was made for the same parameters as in Figure 5.8.

reduce the information density further compared to the raw density; however, they would not alter the trends shown in Figures 5.8, 5.10, and 5.11.

5.2.4 Conclusions

The calculations of the previous sections are summarized in Table 5.1, where the maximum information density $\mathcal{C}$ (in bits/μm^2) for each imaging system is shown along with the values of N_p, b for which it is attained, for two values of material thickness L. We see that the van der Lugt system consistently yields higher information density than the other two systems. However, the optimal values of N_p, b are somewhat

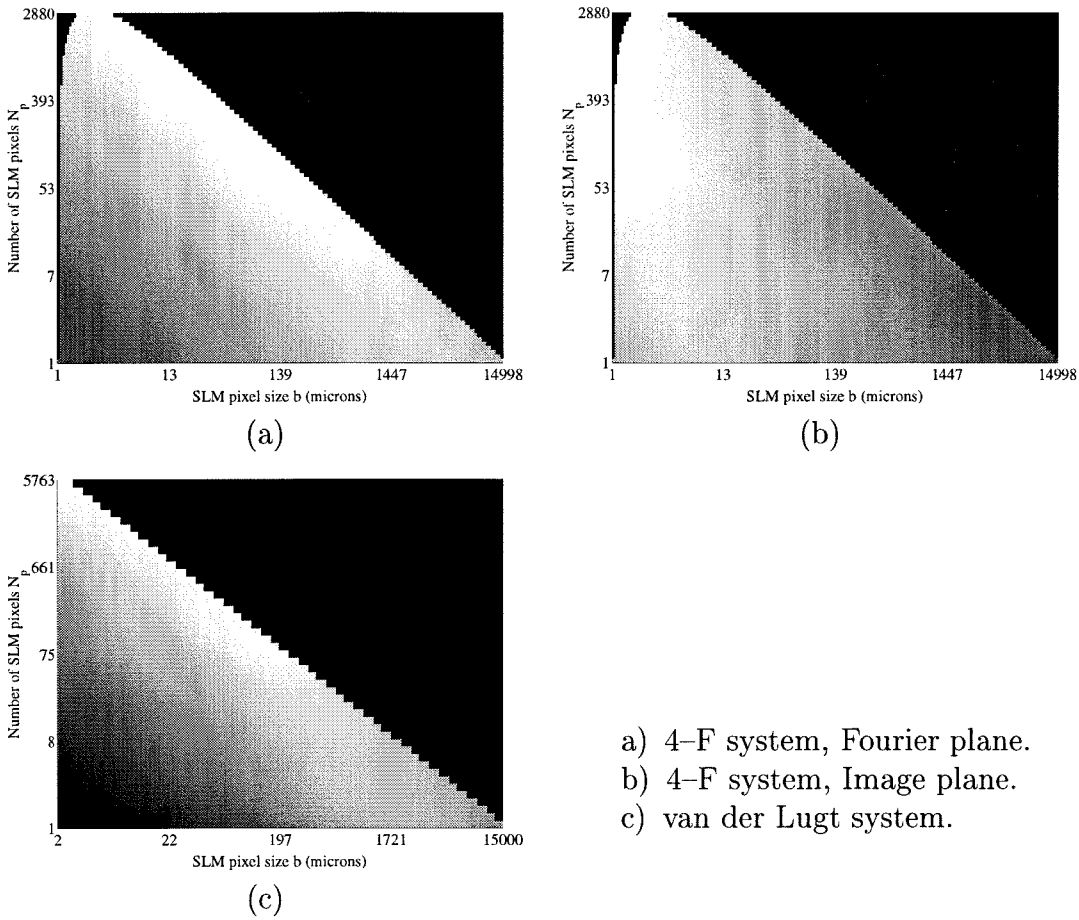


Figure 5.11: Shannon information density versus N_p and b. The brightness in the images is proportional to the information density. The calculation was made for the same parameters as in Figures 5.8 and 5.10.

unrealistic. If we use the more realistic values $N_p = 1,000$, $b \approx 10\mu$m, then the Fourier plane system is superior. Indeed, so far Fourier plane holography with the material slightly displaced from the Fourier plane has been used in all experimental high surface storage density demonstrations [33, 94, 93, 87]. In these experiments, fractal storage was combined with either angle or shift multiplexing to further increase the capacity.

Thin medium, $L = 100\mu$m

	max$\mathcal{C}$(bits/μm^2)	$b_{\text{opt}}(\mu\text{m})$	$N_{p,\text{opt}}$	$\mathcal{C}(b = 10.2\mu\text{m}, N_p = 1000)$
F	16.49	14.29	956	**7.76**
I	13.70	1.43	956	0.39
V	**20.43**	4.91	3657	6.22

Thick medium, $L = 1$mm

	max$\mathcal{C}$(bits/μm^2)	$b_{\text{opt}}(\mu\text{m})$	$N_{p,\text{opt}}$	$\mathcal{C}(b = 10.2\mu\text{m}, N_p = 1000)$
F	116.5	9.16	1399	**64.0**
I	88.7	1.66	1862	3.78
V	**176.1**	2.73	5763	48.0

Table 5.1: Summary of the results for angle multiplexed holographic memories and different imaging systems and thicknesses. "F," "I," "V" stand for Fourier plane, image plane, and van der Lugt system, respectively.

5.3 Wavelength-multiplexed memories

Wavelength-multiplexed holographic memories [37, 38] will undoubtedly become more popular as compact wavelength-tunable sources (laser diodes and vertical-cavity surface-emitting lasers, VCSEL's) in the visible region of the spectrum become widely available. The wavelength selectivity of a general holographic multiplexing geometry is given by (3.17). Most commonly, wavelength multiplexed memories are implemented in the reflection geometry (reference and signal counter-propagating, $\theta_S = 0°$, $\theta_R = 180°$), which yields optimal selectivity[3],

$$\Delta\lambda = \frac{\lambda^2}{2L}, \tag{5.24}$$

where λ is the average wavelength and $\Delta\lambda << \lambda$.

The storage density of wavelength mutiplexing is the same as that of angle multi-

[3]For angle multiplexing, optimal selectivity is obtained in the 90° geometry (also for shift multiplexing, see 3.3.2). Interestingly, the two geometries are symmetric,

$$(\Delta\theta)_{90°} = \frac{\lambda}{2L}, \qquad \frac{(\Delta\lambda)_{\text{refl.}}}{\lambda} = \frac{\lambda}{2L}.$$

plexing (under our assumptions – in [69] the additional volume taken by the reference beam differentiated the two cases); therefore, the equations of section 5.2.1 hold. The number of superimposed wavelength-multiplexed holograms is given by

$$M = \frac{(\Delta\lambda)_{\text{tot}}}{\Delta\lambda}, \tag{5.25}$$

where $(\Delta\lambda)_{\text{tot}}$ is the total tunable range of the laser. Intra-page crosstalk is also obviously the same as for angle multiplexing.

The inter-page crosstalk calculation for wavelength multiplexing in the Fourier geometry was done in [73, 72]. The result is summarized in the following formula:

$$g_j(x,y) = \sum_{m=-M}^{M} f_m(-x,-y)\,\text{sinc}\left((m-j)\frac{2L\Delta\lambda}{\lambda^2}\left(1-\frac{x^2+y^2}{2F^2}\right)\right), \tag{5.26}$$

with the same notation as in section 5.2.2. Using wavelength separation equal to the Bragg selectivity (5.24), we obtain the inter-page SNR

$$\frac{1}{\text{SNR}_{\text{F,inter}}(x,y)} = \sum_{m\neq j} \text{sinc}^2\left((m-j)\left(1-\frac{x^2+y^2}{2F^2}\right)\right). \tag{5.27}$$

This indicates that the SNR becomes worse quadratically, and in a radially symmetric fashion away from the center of the hologram, in contrast to angle multiplexing where the degradation is linear and in one direction only (along the intersection of the plane defined by the reference and signal beams and the surface of the holographic medium).

The derivation of the inter-page crosstalk for image plane wavelength-multiplexed holograms is considerably more complicated than the case of angle-multiplexing; therefore, we will present it in some detail. We begin by Fourier transforming the reconstruction given by the Fourier-plane system of (5.26), obtaining

$$g_j(x',y') = \frac{1}{(\lambda F)^2}\sum_{m=-M}^{M}\iint_{-\infty}^{+\infty} f_m(-x'',-y'')\,\mathcal{B}_{m-j}(x''+x',y''+y')\,\mathrm{d}x''\mathrm{d}y'', \tag{5.28}$$

where

$$\mathcal{B}_n(u,v) \equiv \iint_{-\infty}^{+\infty} \exp\left\{-i2\pi\frac{ux+vy}{\lambda F}\right\} \operatorname{sinc}\left(n\left(1-\frac{x^2+y^2}{2F^2}\right)\right) \mathrm{d}x\mathrm{d}y. \tag{5.29}$$

The following property of $\mathcal{B}$ will prove useful:

$$\iint_{-\infty}^{+\infty} \mathcal{B}_n(u,v)\mathcal{B}_n^*(u+u_1,v+v_1)\mathrm{d}u\mathrm{d}v = \\ = (\lambda F)^2 \iint_{-\infty}^{+\infty} \exp\left\{i2\pi\frac{u_1x+v_1y}{\lambda F}\right\} \operatorname{sinc}^2\left(n\left(1-\frac{x^2+y^2}{2F^2}\right)\right) \mathrm{d}x\mathrm{d}y. \tag{5.30}$$

The proof is straightforward by noting that integrating for u, v yields δ-functions, which then allow two more integrations to reduce the sixfold integral to a double one.

Subsequently, we form the quantity

$$\mathcal{R}_{g_j}(\xi,\eta) = \mathrm{EV}\left\{g_j(x',y')g_j^*(x'-\xi,y'-\eta)\right\} \\ = \frac{1}{(\lambda F)^4} \sum_{m_1=-M}^{M} \sum_{m_2=-M}^{M} \iiiint_{-\infty}^{+\infty} f_{m_1}(x_1'',y_1'')\, f_{m_2}(x_2'',y_2'') \times \\ \mathcal{B}_{m-j}(x_1''+x',y_1''+y')\, \mathcal{B}_{m-j}^*(x_2''+x'-\xi,y_2''+y'-\eta)\, \mathrm{d}x_1''\mathrm{d}y_1''\mathrm{d}x_2''\mathrm{d}y_2''. \tag{5.31}$$

This expression is simplified by using (5.30) and the correlation property (5.17) of the stored holograms. The result, after setting $(\xi,\eta) = (0,0)$ (as explained in section 5.2.2) is

$$\mathcal{R}_{g_j}(0,0) = \mathcal{R}_{f_j}(0,0) + \frac{1}{(\lambda F)^2} \sum_{m\neq j} \iint_{-\infty}^{+\infty} \left\{ \iint_{-\infty}^{+\infty} \exp\left\{i2\pi\frac{\xi'x+\eta'y}{\lambda F}\right\} \operatorname{sinc}(w\xi') \operatorname{sinc}(w\eta') \right. \\ \left. \mathrm{d}\xi'\mathrm{d}\eta' \right\} \operatorname{sinc}^2\left(n\left(1-\frac{x^2+y^2}{2F^2}\right)\right) \mathrm{d}x\mathrm{d}y. \tag{5.32}$$

The internal integral evaluates to $\operatorname{rect}(x/(w\lambda F))\operatorname{rect}(y/(w\lambda F))$. To simplify the calculation further, we approximate the square domain of integration with a disk of

radius $(w\lambda F)/\sqrt{\pi}$ (hence the disk and the square have equal areas). The integrand then becomes radially symmetric, and after some scaling transformations, and using the same definition of image plane SNR as in section 5.2.2, we finally obtain

$$\frac{1}{\mathrm{SNR}_{\mathrm{I,inter}}} = 2\sum_{m\neq j}\int_0^1 \rho\,\mathrm{sinc}^2\left((m-j)\left[1-\frac{(\lambda w\rho)^2}{2\pi}\right]\right)\mathrm{d}\rho. \tag{5.33}$$

Note that setting $w = 0$ earlier would have led to the approximate expression derived by K. Curtis [89, 105] for wavelength multiplexing. However, for our purposes it is important to maintain the dependence of SNR on the pixel size $b \approx 1/w$.

We are now fully equipped to repeat for wavelength multiplexing the calculations for density (which is identical to the density of angle multiplexing), crosstalk, and information density. The results are summarized in Table 5.2.

Thin medium, $L = 100\mu$m

	max$\mathcal{C}$(bits/μm^2)	b_{opt}(μm)	$N_{p,\mathrm{opt}}$	$\mathcal{C}(b = 10.2\mu\mathrm{m}, N_p = 900)$
F	**20.85**	22.93	593	0.40
I	19.70	2.10	335	**18.50**
V	4.95	3.10	4830	1.65

Thick medium, $L = 1$mm

	max$\mathcal{C}$(bits/μm^2)	b_{opt}(μm)	$N_{p,\mathrm{opt}}$	$\mathcal{C}(b = 10.2\mu\mathrm{m}, N_p = 900)$
F	**145.76**	14.78	869	3.92
I	117.27	2.55	1272	**108.13**
V	48.18	2.1	7072	14.01

Table 5.2: Information density of wavelength multiplexed holographic memories for different imaging systems and thicknesses. "F", "I," "V" stand for Fourier plane, image plane, and van der Lugt system, respectively. The parameters used for the calculation were $\lambda = 750$ nm, and $(\Delta\lambda)_{\mathrm{tot}} = 300$ nm, and the rest were the same as in section 5.1.

Comparing to Table 5.1, we see that the conclusions regarding the optimal configuration are quite different for wavelength multiplexing: Fourier plane yields consistently better optimal density, whereas image plane yields consistently higher achievable density with reasonable parameters[4]. The reason is that, at the high wavelength where tunable lasers operate, the penalty to Fourier-type geometries because of crosstalk compared to image-plane geometry is reduced, and also the diffraction spread becomes much more dominant because the pixel size becomes effectively smaller (compared to wavelength). On the other hand, the van der Lugt geometry is penalized because the effect of the diffraction spread doubles since the focus is not in the center of the holographic medium (see Fig. 5.7c). For this reason, Van der Lugt imaging performs worse than the other two types for wavelength-multiplexed memories.

Reflection geometry allows wavelength-multiplexed holograms to be packed most densely, and with minimal crosstalk. However, very expensive polarization optics are also required to minimize optical noise coming from the reference beam leaking into the detector. Alternatively, the transmission and 90° geometries can be used for wavelength multiplexing at the expense of storage density. The above calculations should be modified to account for the decreased density and increased inter-page crosstalk in these sub-optimal geometries.

5.4 Shift-multiplexed memories

The physics of shift multiplexing was described in detail in Chapters 3 and 4. Here we are interested in repeating the imaging system evaluation of sections 5.2 and 5.3 for shift-multiplexed memories.

One feature of shift multiplexing that we must take into account is that the signal is *not* normally incident at the holographic material (as we assumed for angle and wavelength systems), but is rather incident at angle θ. The storage density of shift multiplexing was rigorously calculated in section 4.4. To keep the analysis tractable

[4]Note that the point $(b, N_p) = (10.2\mu\text{m}, 1000)$ is outside the allowable region for this wavelength.

in this section, we will approximate the density by assuming that the area taken up by overlapping shift multiplexed holograms in the Fourier, image, and van der Lugt systems is the area taken by an angle/wavelength multiplexed holograms in the corresponding system (i.e., a_{F}, a_{I}, a_{V} respectively; see section 5.2.1) divided by $\cos\theta$ to account for the signal beam tilt. The number of overlapping holograms is therefore

$$M = \frac{\sqrt{a_{\mathrm{F,I,V}}}}{\delta \cos\theta}, \tag{5.34}$$

where the correct a must be used depending on the geometry, and δ is the shift selectivity, given by (3.67). Another feature of shift multiplexing important for the calculation is that the focal distance z_0 is not arbitrary but is determined by the requirement that the reference and signal beams totally overlap inside the holographic material (see eqs. 4.31 and 4.39 in section 4.4). Here we will use the simple approximation

$$z_0 = \frac{\sqrt{a_{\mathrm{F,I,V}}}}{(F/\#)\cos\theta}, \tag{5.35}$$

where $(F/\#)$ is the F-number of the lens used at the reference arm. We will always assume that the reference and signal lenses have the same $(F/\#)$.

The interpage crosstalk calculation for shift multiplexing is straightforward. For Fourier plane holograms we use directly (4.4). For image plane holograms the calculation is similar to the angle multiplexing case, and yields

$$\frac{1}{\mathrm{SNR}_{\mathrm{I,inter}}} = \sum_{m \neq j} \int_{-1}^{+1} (1 - |\xi'|)\, \mathrm{sinc}^2\left(\lambda w (m-j)\frac{\xi'}{\sin\theta}\right) \exp\{i2\pi(m-j)\xi'\}\, \mathrm{d}\xi'. \tag{5.36}$$

The storage density, crosstalk and information density comparisons of different imaging systems for shift multiplexing are now straightforward. The results are summarized in Table 5.3.

Concluding, we observe that overall the shift multiplexing method yields higher capacities than angle or wavelength, because it allows more efficient usage of the

Thin medium, $L = 100\mu$m

	max$\mathcal{C}$(bits/μm^2)	$b_{\text{opt}}(\mu\text{m})$	$N_{p,\text{opt}}$	$\mathcal{C}(b = 10.2\mu\text{m}, N_p = 1000)$
F	34.22	14.21	956	17.87
I	27.26	1.51	1272	0.75
V	**43.95**	4.58	3048	**20.85**

Thick medium, $L = 1$mm

	max$\mathcal{C}$(bits/μm^2)	$b_{\text{opt}}(\mu\text{m})$	$N_{p,\text{opt}}$	$\mathcal{C}(b = 10.2\mu\text{m}, N_p = 1000)$
F	208.0	9.16	1399	120.7
I	165.3	1.66	1862	7.02
V	**371.9**	1253	73	**194.5**

Table 5.3: Shannon information density of shift multiplexed holographic memories for different imaging systems and thicknesses. "F," "I," "V" stand for Fourier plane, image plane, and van der Lugt system, respectively. All parameters are the same as in section 5.1.

apertures, and also has reduced crosstalk. The van der Lugt system is preferrable in all cases. One possible limitation of shift multiplexing that we did not take into account is the leakage of the wide-angle reference beam into the detector during reconstruction thereby increasing the noise and decreasing the capacity. In practice so far this was not a problem [87]; however, pushing the system to the limits as we attempt to do here might make this consideration significant.

Chapter 6 Issues in holographic memory design

In this chapter we discuss practical issues in holographic memory design. In section 6.1 we describe two methods for non-volatile storage in photorefractives, namely electrical fixing (section 6.1.1) and periodic refreshing with the aid of phase conjugation (section 6.1.2). In section 6.2 we elaborate further on a compact architecture for dynamic holographic memories and design a Terabit holographic memory that can fit in a volume quite smaller than 1 m^3. Finally, in section 6.3 we begin the treatment of the issue of access to the database by presenting various architectures for associative storage. The topic of memory design and interface will be treated fully in the next chapter.

6.1 Volatility in photorefractive holographic memories

Volatility in photorefractive storage is a major concern, in the same fashion as in silicon DRAM memories. Erasure in photorefractive materials is well understood [35, 107], and is closely related to the recording process itself. We will give a brief description here before we discuss in more detail two techniques, namely electrical fixing and periodic refreshing for overcoming the volatility problem.

The photorefractive effect is based on photo-induced band-transport of electrons, as shown in Fig. 6.1. The photorefractive crystal is illuminated by a sinusoidal light pattern $I(x)$. Electrons that happen to lie in energy levels between the valence band and the conduction band (i.e., electrons that belong to deep traps) and were initially located at illuminated regions are excited to the conduction band, and move freely

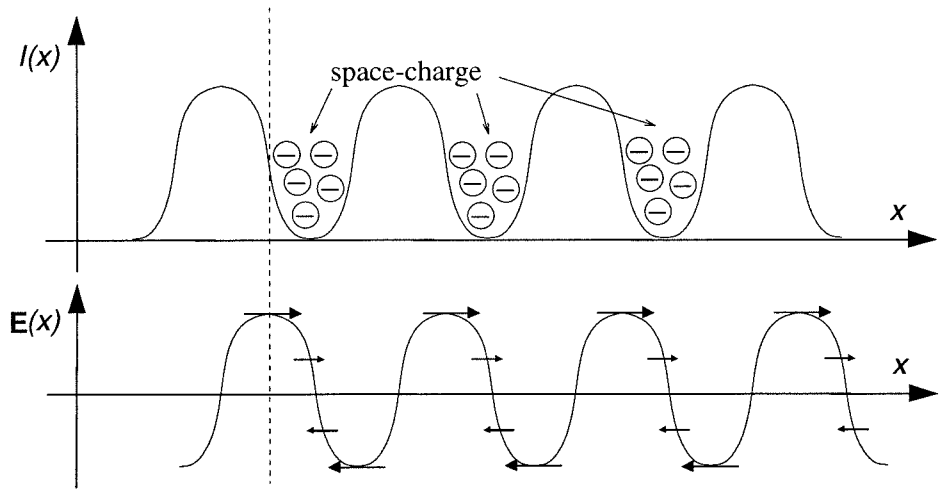


Figure 6.1: Simple diagram explaining the photorefractive effect in the diffusion-dominated case.

to the dark regions, because of diffusion. There they recombine with ions at acceptor sites. (For simplicity we assume that there are no external electric fields or photovoltaic fields, i.e., there is no drift.) When steady-state is reached, excess electrons will have accumulated in the dark regions, generating a space-charge field $\mathbf{E}_{\text{sc}}(x)$ out of phase by $\pi/2$ with respect to the illumination. In turn, $\mathbf{E}_{\text{sc}}$ locally modulates the refractive index of the crystal via the electro-optic effect, generating a phase hologram. When drift is present, the recording mechanism is very similar, except the phase-shift between $\mathbf{E}_{\text{sc}}(x)$ and $I(x)$ is different than $\pi/2$. The quantitative treatment of this process is based on the band-transport equations [35, 90, 107, 108].

The phase hologram would be stable if the space-charge pattern developed during the recording phase could be maintained. Unfortunately, this is not true even in the dark, because some of the recombined electrons are thermally excited back to the conduction band, and the space-charge field decays. The dark-storage lifetime depends on the doping species and concentration, and can be as low as a few seconds (typically in $BaTiO_3$) or as high as a few weeks (in lightly doped $LiNbO_3$). When the hologram is illuminated, the thermal excitation of recombined electrons is enhanced by photoexcitation, and the decay is stronger. Therefore, photorefractive holograms are optically erasable. This is convenient for the HRAM application. However, two situations exist when optical erasure is undesirable: hologram readout, and recording

of multiple holograms.

To record multiple holograms, one must determine appropriate recording times, i.e., follow an "exposure schedule," [51, 91, 109] so that, after each exposure, all the stored holograms have equal diffraction efficiency[1]. As a result of the simultaneous recording and erasure of previously recorded holograms, the diffraction efficiency after superimposing M holograms in the same location of a photorefractive medium varies with M as [110]

$$\eta = \left(\frac{M/\#}{M}\right)^2 \tag{6.1}$$

where the $M/\#$ (M-number) is a system metric, determined by the material constants (refractive index, absorption coefficient, dopant concentrations, electron mobilities), the beam intensities, and the optical system used for the measurement. (For the calculation of $(M/\#)$ in the case of photorefractive shift-multiplexed memory, see section 4.2).

Several methods have been devised against erasure in the cases of readout and long-term dark storage in photorefractive materials. Historically, the first attempt was to fix the holograms by generating gratings that are not optically or thermally sensitive. One method is to induce ionic gratings by heating up the crystal [95], since then the mobility of hydrogen ions increases. Alternatively, ferroelectric gratings are formed when a negative (with respect to the dielectric polarization) electrical pulse is applied [96], because the ferroelectric domains are locally destabilized. The topic of electrical fixing is discussed in more detail in section 6.1.1.

Instead of fixing the holograms thermally or electrically, another idea is to try to sustain them as long as possible by reading out using a wavelength where the material absorption is low. For example, Fe-doped $LiNbO_3$ has high absorption in the green and blue wavelengths, but is relatively insensitive to red. This two-lambda technique [82] involves trade-offs in the achievable density, because multiple scalar

[1]The complete discussion of exposure schedule for shift-multiplexed photorefractive memories was given in section 4.2.

volume holograms cannot be entirely Bragg-matched in a wavelength other than that used for recording [99, 111], except at the expense of allowing Bragg degeneracies. Section 4.5 provides a full account of the two-lambda method in shift-multiplexed memories.

Holograms may be recorded in Fe-doped $LiNbO_3$ with infrared light by using sensitizing high-intensity green pulses [112]. Commonly, $LiNbO_3$ materials possess shallow traps, i.e., energy levels very close to the conduction band. Electrons from the deep traps are "pushed" to those levels by the sensitization pulses, and can easily be excited to the neighboring conduction band by the low-energy low-absorption infrared light (two-photon process). Since the hologram is recorded and reconstructed in the same long wavelength, the two-lambda readout limitations are eliminated.

Finally, a method that requires neither a complicating fixing process nor multiple wavelengths is periodic copying [98], as in silicon dynamic random access memories. Holograms that coexist in the holographic memory are periodically read-out, fed-back and re-recorded continuously, each for time long enough to maintain a stable system. This technique requires very robust alignment, since any slight deviation in the direction of the fed-back signal results in the opposite effect of efficiently erasing the stored hologram. A natural way of side-stepping the alignment issue is by replacing the resonator with an optoelectronic latch, which is used to detect and re-record the holograms. This idea led to the development of a lens-less compact holographic memory module design and is discussed more fully in sections 6.1.2 and 6.2.

6.1.1 Electrical fixing

Electrical fixing was developed in the early 70's by Micheron and Bismuth as an alternative and hopefully more practical method for non-volatile storage. The method is based on the reversibility of ferroelectric domains in photorefractive crystals. It was first applied in iron-doped barium titanate [96] (Fe:$BaTiO_3$) and then to strontium barium niobate [113] (SBN:x, $Sr_xBa_{1-x}Nb_2O_6$).

Electrical fixing of a single hologram is very simple to implement: during the

poling process of the ferroelectric crystal, the optical $\hat{\mathbf{c}}$-axis is defined, and is parallel to the spontaneous polarization axis, i.e., the direction in which the ferroelectric domains point. In the simplest version, the holograms are recorded without any applied field; after recording is complete, the hologram is fixed by applying a negative, i.e., antiparallel to the $\hat{\mathbf{c}}$-axis electric field $\mathbf{E}_a$. Micheron and Bismuth observed [113] in the SBN specimen used in their experiments the following effects: the field had little effect if $|\mathbf{E}_a| < 700$ V; for 700 V$< |\mathbf{E}_a| <$1000 V the hologram was not optically erasable and the diffraction efficiency η was enhanced; for $|\mathbf{E}_a| >$1000 V, η dropped as $|\mathbf{E}_a|$ increased. By applying a positive electric field to the crystal, the hologram became optically erasable again.

Recently, this fixing process in SBN was revisited and studied extensively [97, 114, 115, 116, 117], with emphasis on the application to practical angularly multiplexed memories. Our current understanding of the fixing process is based on the switching of ferroelectric domains when electric fields are applied [113, 114]. We will describe the physics of fixing based on a typical fixing experiment [97] shown in Figure 6.2.

Consider a photorefractive crystal like SBN, with relatively low coercive field. During holographic recording, the photorefractive space-charge field $\mathbf{E}_{sc}$ builds up and reaches steady state inside the crystal. After recording is complete, the negative pulse $\mathbf{E}_a$ is applied and achieves the following effects: In locations where the total field $|\mathbf{E}_{sc} + \mathbf{E}_a|$ exceeds the coercive field $\mathbf{E}_{co}$, the domains are reversed under the electric force applied to them. If $\mathbf{E}_a$ is strong enough, the switching condition is fulfilled only in regions with negative space-charge field. Thus the space-charge grating field triggers the generation of a polarization grating inside the crystal. The additional electric field generated by the reversed electric dipoles screens the photorefractive space charge field resulting in re-distribution of the space-charge (point C in Fig. 6.2).

Under constant illumination, excess charges that are not trapped by the polarization field are re-distributed. During this time, the electric field experiences a sign reversal (point D in Fig. 6.2). The total electric field in equilibrium $\mathbf{E}_{tot}$ (point F in Fig. 6.2) can be decomposed into two distinct contributions, one due to the polarization grating, and one due to the re-distributed free carriers (electrons). Both

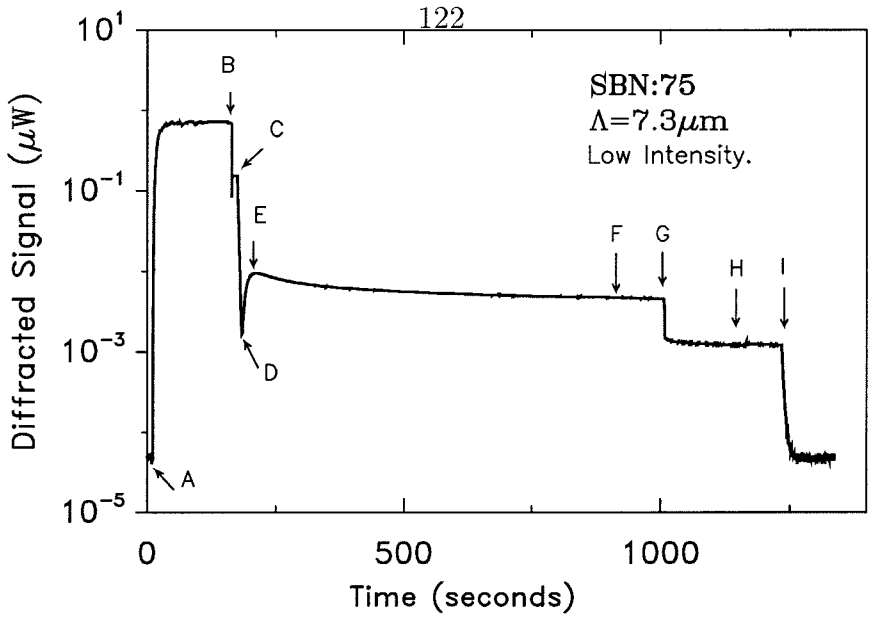


Figure 6.2: A typical fixing-revealing experiment. A: recording begins; B: writing beams are blocked, and negative voltage pulse is applied; C: optical erasure with non-Bragg-matched beam begins; D: phase reversal in the optical field; E: fixed grating reaches peak value; F: fixed grating reaches steady state; G: positive pulse is applied; H: revealed (compensating) grating; I: optical erasure.

components are proportional to the amplitude P_s of the polarization grating [114].

When a positive pulse is applied to the crystal with the fixed hologram (point G in Fig. 6.2), the ferroelectric domains are switched back in the direction of the $\hat{\mathbf{c}}$-axis (if the applied field exceeds the coercive field) and thus the ferroelectric domain grating is destroyed. This results in the revealing of the compensating charges and enhanced diffraction efficiency, since screening is cancelled. The revealed hologram is electronic; therefore, it is optically erasable (point I in Fig. 6.2).

The explanation of the fixing effect given so far agrees with the observations of Micheron and Bismuth [113] and Qiao et al. [97]. The fixing efficiency depends strongly on the grating spacing Λ. This is expected, since the efficiency of switching should depend on the relative sizes of the ferroelectric domains and the space-charge field period. The ratio of the efficiency of the compensating grating to the efficiency of the fixed grating increases as $(\Lambda/\Lambda_D)^4$, where Λ_D is the Debye length. This dependence is predicted theoretically from the charge transport equations, and is also

	Relative diffraction efficiency	
	fixed hologram	revealed hologram
$\mathbf{E}_a = -2.2\hat{\mathbf{c}}$ kV/cm	$0.74\eta_0$	$0.06\eta_0$
$\mathbf{E}_a = -3.0\hat{\mathbf{c}}$ kV/cm	$< 1 \times 10^{-3}\eta_0$	$0.40\eta_0$

Table 6.1: Fixing efficiency versus fixing pulse amplitude. η_0 denotes the diffraction efficiency of the hologram before fixing ($\sim$ 30% in most experiments) for grating spacing $\Lambda = 10\mu$m. The data show that there is a sharp threshold at approximately 2.6kV/cm; below, fixing is very inefficient, as shown by the low-efficiency revealed hologram. Above, the opposite happens, since a single positive pulse is sufficient to reveal significant portion of the original hologram. After 12 pulses, as much as $0.5\eta_0$ was revealed in the same experiment.

verified experimentally independent of the recording and erasing intensities [114].

The diffraction efficiency of the fixed hologram in steady state for small grating spacings Λ increases [115] as Λ^p with $p = 1.3 \sim 2.1$ (p decreases with recording/erasing intensity) and peaks at $\Lambda \approx 7\mu$m. Adding to this effect the Λ^4 dependence of the relative compensating-to-fixed grating efficiency, we find that the compensating grating increases like $\Lambda^{5.3\sim6.1}$, and this is more or less verified experimentally [115]. The revealed grating itself has maximum diffraction efficiency at $\Lambda \approx 11.2\mu$m.

These observations depend critically on the dynamics of the formation of the ferroelectric domain grating. Recently, it was shown that the fixing/revealing fields also have a strong effect on the fixing efficiency [116]. Applying a small amplitude negative pulse in general has the effect of creating a weak domain grating. In that case, the fixed hologram has appreciable efficiency, but the revealed hologram is weak. On the contrary, a high-amplitude negative pulse creates a strong domain grating, which results in a very weakly diffracting fixed grating (since all or most of the space-charges are screened through compensating the domain grating), and a very strong revealed grating (up to 50% of the diffraction efficiency of the grating before fixing). These results are summarized in the data of Table 6.1.

The understanding of the physics of electrical fixing made it possible to fix angularly multiplexed holograms in SBN:75 [116]. The multiplexing setup is very similar to

that of Fig. 1.1. The angle θ between the reference and signal beams has to be in the order of a few degrees for efficient fixing to occur (the optimal $\Lambda = 7\mu$m corresponds to $\theta = 1.6°$). This requirement leads to the following trade-offs: *(1)* the number of holograms increases as θ is made larger, because the angle selectivity $\Delta\theta$ goes approximately like $\Delta\theta \approx \lambda/n\theta$; *(2)* the signal-to-noise ratio improves as θ increases, because the reconstruction is less affected by scatter noise coming from the reference and crosstalk noise coming from partial reconstruction of overlapping holograms; *(3)* the individual capacity of the holograms increases with θ because the available angular bandwidth becomes larger.

Very recently, 1,000 holograms were multiplexed and electrically fixed [117, 118] in SBN:75 (0.02% Ce-doped, thickness 1 cm) using the method described above. The holograms were arranged in 5 fractal rows [40, 31], each containing 200 angle-multiplexed holograms. The signal-reference interbeam angle was chosen $\theta \approx 3°$. The holograms were recorded initially using the usual recording schedule [91], and a negative fixing pulse was applied immediately afterwards. The diffraction efficiency of the individual fixed holograms was below noise level. After revealing, the average diffraction efficiency was $\approx 0.005\%$ with 80% uniformity. No degradation was observed after several fixing-revealing cycles.

We described in detail one method of domain fixing in ferroelectric materials. Variants of this idea also exist. For instance, applying a negative field *during* recording has been reported in both SBN [97] and barium titanate [119]. In the former case, the diffraction efficiency of the fixed grating was found to be higher than if the fixing pulse were applied after recording. It is even possible to obtain fixed holograms in SBN:75 by applying the negative pulse *before* recording, i.e., by preliminary partial de-poling of the crystal [117].

One potential limitation in the long term stability of fixed domain gratings is the electrostatic energy stored in the domain walls, which occurs because of the non-zero polarization gradient $\nabla \cdot \mathbf{P}$ (the spontaneous polarization is along the direction of the grating vector). By recording gratings perpendicular to the $\hat{\mathbf{c}}$-axis while applying an electric field anti-parallel to the $\hat{\mathbf{c}}$-axis, Horowitz et al. demonstrated a new domain

fixing technique based on the screening field [120]. The screening field cooperates with the external field in regions of high light intensity to cause domain reversal and works even though the domains are perpendicular to the grating vector. This method can be applied to fix patterns without the presence of a reference beam, and also to weakly photorefractive crystals. According to the authors, features as small as 1μm can be fixed. To our knowledge, no multiplexing results using this method have been reported to-date.

6.1.2 Periodic refreshing

The decay of holograms in photorefractive materials during illumination or even in the dark is reminiscent of the effect of leakage current through the output capacitors in silicon memories. In the latter, the data decay problem is solved with a periodic read/re-write sequence, called refreshing. The periodic refreshing idea may also be applied to holographic memories and several refreshing architectures have been devised [121, 122, 123, 124]. The general refreshing idea is to incorporate the reconstruction into a feedback loop; if the fed back signal is strong enough and of sufficient fidelity, then it can be recorded on top of already existing holograms in order to reinforce them. Usually distortion-free feedback is obtained by the well-known technique of phase conjugation [125, 126, 127].

In a phase-conjugate system, distortion correction is obtained as follows [128]: suppose that a hologram is recorded by a reference beam R and a signal (object) beam $Se^{i\phi}$ (Fig. 6.3a), where S is the actual signal and ϕ is the phase aberration introduced by the beam propagation[2] (including Fresnel diffraction). The interference pattern is expressed as $\left|R + Se^{i\phi}\right|^2$. When R is used for reconstruction (Fig. 6.3b), $Se^{i\phi}$ is obtained on the signal axis as a continuation of the signal beam, carrying over all the phase aberrations introduced in the signal path during recording. Therefore, the forward reconstruction is distorted and needs correction by using the appropriate imaging optics. If, however, the phase-conjugated reference R^* is used for reconstruc-

[2]It is usually safe to ignore the effects of absorption on the phase-conjugation process.

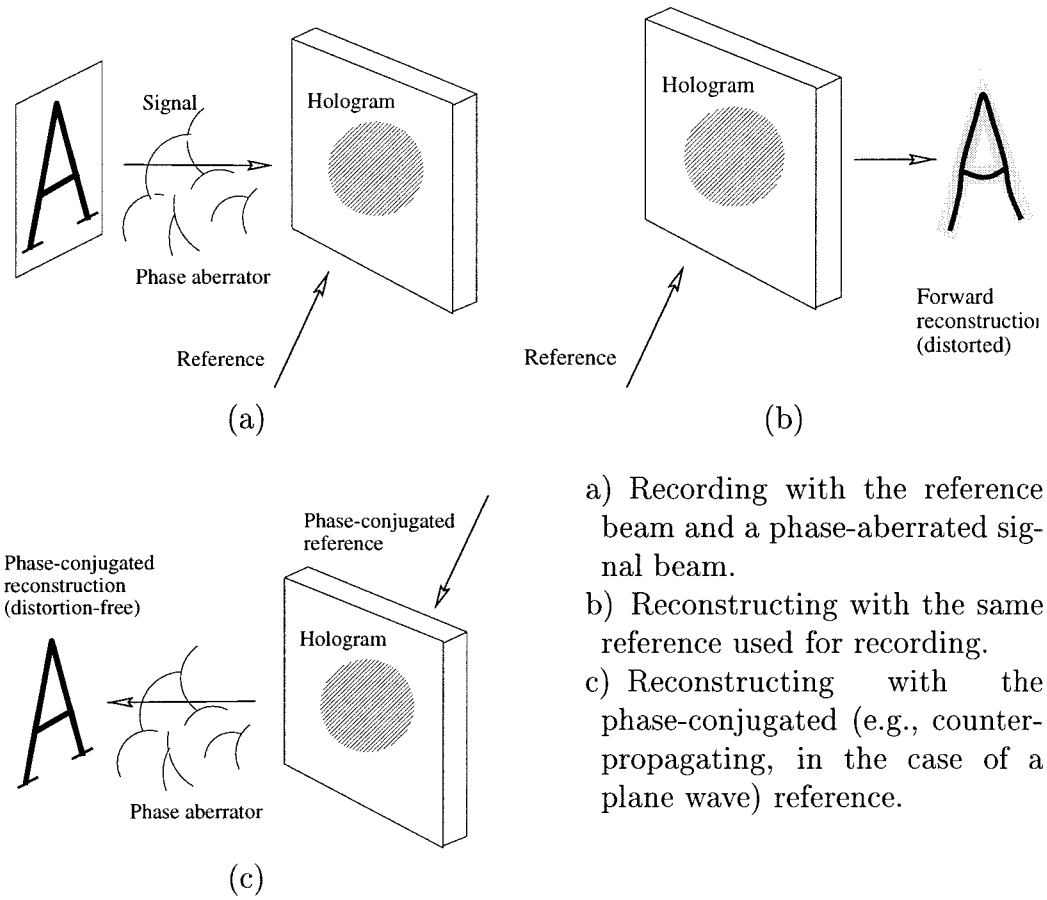


Figure 6.3: Obtaining distortion-free hologram reconstruction with a phase-conjugated reference.

tion (Fig. 6.3c), the on-axis reconstruction produced by the hologram contains the term $S^*e^{-i\phi}$. The reciprocal aberrator introduces an additional phase term $e^{i\phi}$ to the counterpropagating reconstruction, and, as a result, a distortion-free intensity image $|S|^2$ is obtained at the location of the original signal.

More recently, the phase-conjugate reconstruction method has been used in the design of a compact refreshable dynamic holographic memory with liquid-crystal optoelectronic interface [129, 130]. In this design, an optoelectronic circuit, the Dynamic Hologram Refresher (DHR) participates in the feedback loop by detecting, thresholding, and re-recording holograms periodically. Phase conjugation contributes to the compactness of the system because it eliminates the need for imaging optics in the

design. The relevant design and optimization issues are presented in detail in the next section.

6.2 Compact design of a Terabit Random-Access Memory

Compared with commercial optical memories (e.g., digital video disks), holographic memories offer provably at least equal capacity [33], and potentially both higher capacity and faster access time, by several orders of magnitude. However, holographic storage also poses more stringent requirements on the quality of the recording and readout optical systems, and may take more space because of the bulky laser, optics, and vibration isolation equipment. It is therefore desirable to design holographic memories of physical size much smaller than currently available. Ideally, maximum robustness and compactness are achieved if the optical elements are placed close enough so that they can be glued to each other. Miniaturized optical sources (e.g., vertical-cavity surface-emitting lasers, VCSEL's) and components (beam-splitters, mirrors, waveplates) are readily available today, and optoelectronic technology is mature enough to provide reliable, high-performance integrated interfaces such as Spatial Light Modulators [131, 132, 133] (SLM's) and detector arrays. However, the imaging requirements [103, 44, 134] for high-capacity holographic storage usually force the designer to use expensive and bulky aberration-corrected lenses [33]. The phase conjugation technique (section 6.1.2, [125, 126, 127]) allows wave-front distortion recovery without imaging in a properly designed architecture. It has been used, for example, to design a read-only holographic memory with a conventional detector interface [135].

In this section we are interested in compact dynamic refreshable holographic memories, and in particular in the implementation of a Tbit holographic memory with volume comparable to that of a desktop personal computer. This corresponds to system volume density of several Tbits/m^3. To get an idea of the orders of magni-

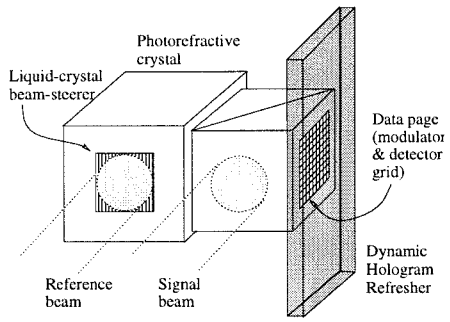
tude involved, consider that the volume density of the Merriam-Webster dictionary is 60 Gbits/m^3, and the Millikan Memorial Library at Caltech stores approximately 3 Tbits of information in 450,000 volumes. Therefore, our proposed system is capable of storing the contents of the entire library in a box smaller than a cubic meter!

We begin with the design of the basic compact memory module and some related experimental results in section 6.2.1. Subsequently we consider some of the basic issues in holographic memory design for our particular architecture, namely multiplexing method (section 6.2.2), system density optimization (sections 6.2.3, 6.2.5), and the trade-off between power, noise, error rate, and access time (section 6.2.4). Discussion, improvements, and some general comments are given in section 6.2.6.

6.2.1 Compact dynamic holographic memory architecture

The basic module used in dynamic compact holographic memories is shown in Figure 6.4. It consists of a photorefractive crystal, which acts as a re-writable holographic material, a polarizing beam-splitter, and an optoelectronic integrated circuit, the Dynamic Hologram Refresher [129, 130, 136] (DHR). The plane-wave reference beam is directly incident to the crystal through a liquid-crystal beam-steerer [137] for angle multiplexing. The signal beam is first deflected by the beam-splitter towards the optically active surface of the DHR die, where the signal information intensity-modulates the waveform. The modulated signal beam is reflected back to the photorefractive crystal where it records a hologram with the reference beam. The phase-conjugate method (see section 6.1.2) is employed for reconstruction. A counter-propagating reference beam passing through an identical beam-steerer (neither is visible in Fig. 6.4a) reconstructs the phase-conjugated signal which counter-propagates towards the DHR.

The counter-propagating beam may be provided by a mirror coating at the back face of the photorefractive crystal; in that case, the input beam must be steered at the complement of the desired readout angle. Yet another way to provide the phase-conjugate beam is by using a self-pumped phbase conjugator in a separate crystal [130]. This solution is very elegant but it adds to the volume of the module



(a)

Pixel size	132μm×211μm
Active pixel area: (modulator)	49μm×49μm
(detector)	10μm×10μm
Number of pixels	20 × 24
Equivalent noise power	≤100 fWatts
Contrast ratio	18:1

(b)

Figure 6.4: (a) Basic module of a compact holographic memory. (b) Operation characteristics of the Dynamic Hologram Refresher.

and is expensive. In our experimental demonstration we used a Sagnac interferometer configuration to provide the phase-conjugated reference. In the theoretical derivations and volume optimization (section 6.2.3), we will assume that the mirror solution is used.

The surface of the DHR is organized as a grid of pixels, each containing a metal pad and a phototransistor. A layer of hybrid-aligned nematic (HAN) liquid crystal is sandwiched between the silicon die and a transparent grounding electrode coated with indium-tin oxide (ITO). When a voltage is applied to the metal pad of a particular pixel, the phase of light incident to and reflected by the metal pad is modulated due to the electro-optic effect in the intervening liquid crystal [133]. Thus, with the aid of a polarizer, this device acts as an intensity modulator for each individual pixel.

The operating characteristics of the DHR are given in Figure 6.4b. The DHR functionality is threefold:

(i) In the hologram recording phase, the DHR modulates incident light and reflects the modulated beam onto the photorefractive crystal, where a hologram is recorded by interfering with the reference.

(ii) In the hologram reconstruction phase, the diffracted beam obtained by the phase-conjugate method counterpropagates back onto the phototransistors, where the reconstructed pattern is detected by the DHR. Note that since the detec-

tors are not collocated with the metal pad modulators, a truly phase-conjugated beam cannot be detected in the way we described so far. This problem is solved by introducing a small tilt in the reference beam along the degenerate direction, and has been proven to work in practice [129].

(iii) In the hologram refreshing phase, the two previous operations are combined inside the DHR without external influence as follows: first the hologram is reconstructed by the phase-conjugated reference, detected as described in (ii), and latched in the internal memory of the DHR; then the detected pattern is transferred to the modulators and is used as signal to re-record a hologram with the forward reference thus reinforcing the original hologram.

The capacity of the basic module described so far is of the order of 1 Gbit, assuming that roughly 1,000 holograms can be stored in the photorefractive crystal, and each one contains approximately 1 Mbit (1000×1000 pixels). This goal is somewhat optimistic given the current state-of-the-art in integrated optoelectronic technology and holographic materials; however, it is not far from being realizable [136]. We will pick up on this point in sections 6.2.3 and 6.2.4.

The experimental setup used[3] for testing the basic dynamic memory module is shown in Figure 6.5. Polarizing beam splitter PBS1 splits the input beam into two arms, the reference and signal. The reference is directed into the Sagnac interferometer formed by PBS2 and mirrors M1, M2, M3. When the interferometer is aligned, the two counterpropagating beams are phase-conjugated. The signal, after passing through PBS3, is incident on the DHR where it gets modulated as described in section 6.2.1 before being reflected back towards the photorefractive crystal PRC, 30°-cut $BaTiO_3$ for this experiment. The arm of the reference beam reflected by PBS2 is used for recording, while the transmitted beam is used for phase-conjugate reconstruction. Mechanical rotation of the crystal in the plane of the figure is used to implement

[3]Preliminary experiments with this architecture were performed by the author with Jean-Jacques P. Drolet, who also designed and fabricated the DHR chip. These experiments were published in [130]. The more conclusive results presented here and in [129] were obtained by Jean-Jacques P. Drolet and Ernest Chuang. Currently working on this project are Ernest Chuang, Xu Wang, and Wenhai Liu.

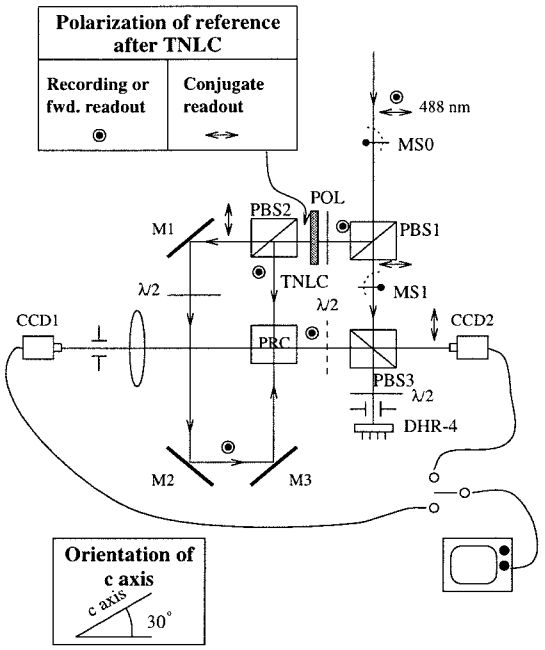


Figure 6.5: Experimental setup for testing the DHR module.

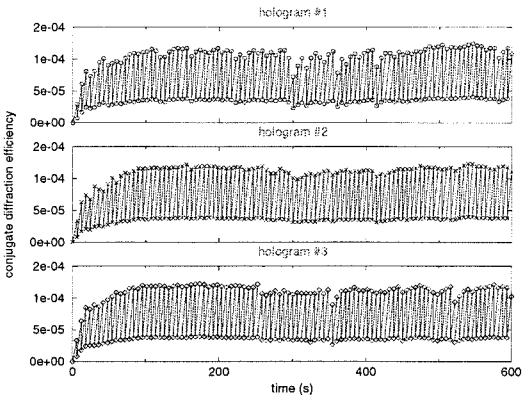


Figure 6.6: Sustainment of three holograms in the phase-conjugate reconstruction geometry using the DHR chip.

angle multiplexing. In addition to detecting the reconstruction on the DHR, we also used two CCD cameras to observe both the forward and the phase-conjugated reconstruction for characterization purposes.

The operation of the DHR as refresher with multiple holograms is shown in Figure 6.6. In this experiment we stored three holograms and used the refreshing method described above to sustain and amplify them to saturation for 100 cycles. No errors were observed in any of the reconstructions. The probability of error, estimated from the pixel intensity statistics, was of the order of 10^{-3}. Sample images obtained by the DHR from the experimental setup of Figure 6.5 are given in Figure 6.7. The calculated error probabilities for images a, b, c, d (please see caption) were 1.1×10^{-4}, 2.2×10^{-3}, 6.9×10^{-4}, 1.0×10^{-3} respectively. This shows not only that the phase-conjugated reconstruction is more reliable than the forward reconstruction (this was expected because of the self-correcting properties of phase-conjugation), but also that the deterioration resulting from multiple refreshing cycles is minimal. As of the writing of this thesis, up to 25 sustained holograms have been demonstrated with similar performance.

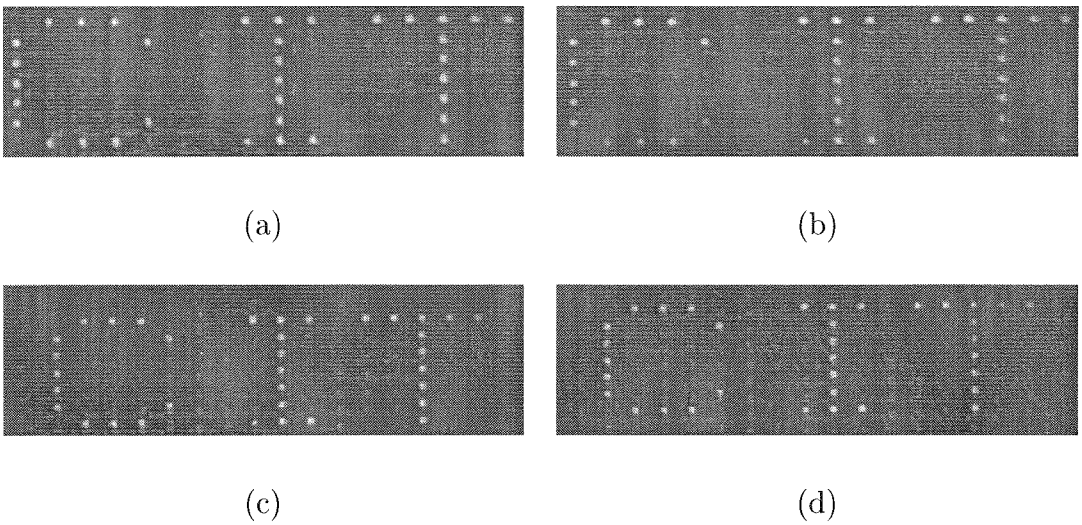


Figure 6.7: (a) DHR display; (b) reconstruction obtained with the forward reference, and conjugated reconstructions after (c) 1 and (d) 50 refreshing cycles.

6.2.2 Selection of multiplexing method

Angle multiplexing (along with fractal and peristrophic) has been widely used in high-capacity holographic memory demonstrations [22, 31, 33, 34]; therefore, it is the first candidate that comes to mind for the compact memory as well. The main challenge in the compact angle-multiplexing implementation is beam deflection. Since mechanical or acousto-optical deflectors are unacceptable, liquid-crystal beam-steerers [137] are one solution. Reflective liquid-crystal beam-steerers operating in the wavelength range of the final compact memory (670 nm, see section 6.2.3) are currently under development by Xu Wang in the Psaltis Laboratory. Alternative solutions utilizing VCSEL (or diode) arrays for angle multiplexing have been proposed[4], but will not be described in this thesis.

Multiplexing methods other than angle alleviate the need for beam deflection; however, they suffer from other disadvantages. More specifically:

1. wavelength multiplexing requires a tunable source over a broad spectral range in order to achieve high capacity;

[4]Ernest Chuang, private communication

2. phase-code multiplexing requires a Fourier-transforming system which would increase the volume; the Fourier transform may be omitted, but then phase-conjugation cannot be achieved with a simple mirror (the only way is a self-pumped phase-conjugator which would worsen the volume and optical power requirements of the system);
3. shift multiplexing so far was described as requiring mechanical translation; however, in section 6.2.5 we will describe a compact implementation with a VCSEL (or diode) array, which competes closely with the angle multiplexing setup, because it yields comparable density without requiring the beam-steerer.

We will consider in detail angle multiplexing with a reflective beam-steerer as a potentially practical solution. Before concluding this section, we will describe the shift-multiplexed implementation.

6.2.3 System volume optimization

One of the advantages of the modular architecture described in the previous sections is that several of these modules may be combined in order to achieve high capacity. For example, if each module holds 1 Gbit (see section 6.2.1), then one may arrange 1,000 modules in a $10 \times 10 \times 10$ grid to obtain a Tbit memory. This calculation, however, is deceptively simple. During the high-capacity memory construction, several other issues must be taken into consideration, e.g., the location, number, and distribution of laser sources, power dissipation, mechanical stability, and system cost [136]. In this section we do a detailed optimization of the volume of the Tbit system versus certain design parameters, taking into consideration constraints imposed by technology.

The modular architecture that we consider in this section is shown in Figure 6.8. The basic module is identical to the angle-multiplexed module of Figure 6.4a, except with reflective liquid-crystal beam-steerers. The optics have been arranged carefully so that the spatial bandwidth is sufficient, i.e., that no significant amount of light is lost from any first-order diffraction beam, at the same time without wasting any extra space. Phase conjugation is achieved by means of a mirror coated on the crystal sur-

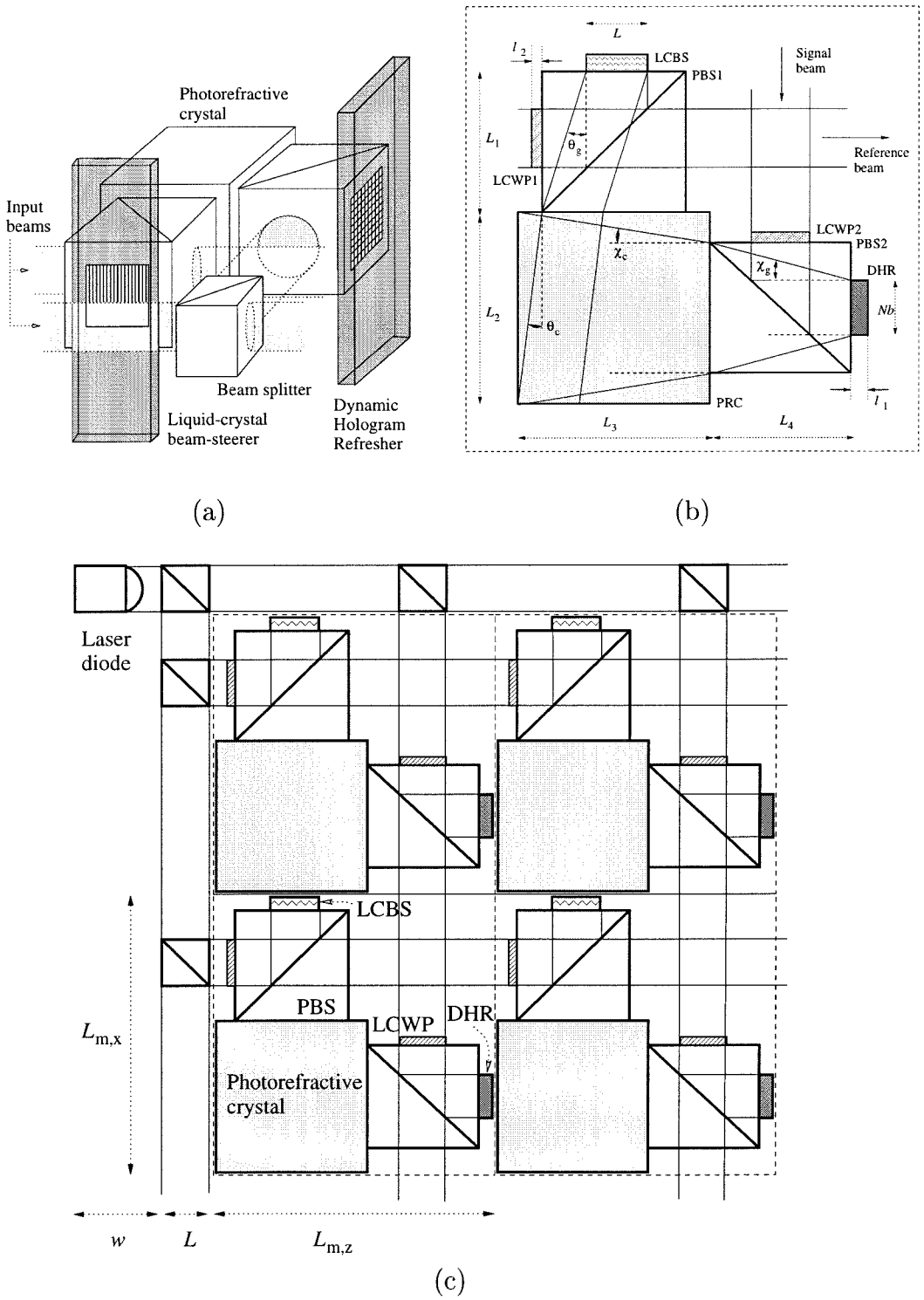


Figure 6.8: Schematic of the modular architecture. (a) Three-dimensional view of the basic module; (b) top view of the basic module; (c) arrangment of basic modules into a $G_x \times G_z$ (here 2×2) grid.

face (not shown in the figures). Also, several waveplates (some of them implemented as liquid-crystal cells so that the phase rotation can be externally controlled) are in place to ensure the appropriate polarizations.

The active DHR surface contains $N \times N$ pixels, each one of size $b \times b$. The fill factor (which takes into account supporting circuitry surrounding the pixel metal pad) is ϕ; therefore, the entire die area is $N^2 b^2/\phi^2$. The height of the DHR chip (perpendicular to the page level in Fig. 6.8b) is h. The beam diameter is L, and because of the special 90°-geometry arrangement, it is also equal to the hologram thickness, needed for the purpose of determining the Bragg selectivity. The sizes of the optical elements are $L_1 \times L_1$, $L_2 \times L_3$, and $L_4 \times L_4$ for polarizing beam splitter PBS1, photorefractive crystal, and PBS2, respectively. Additional lengths needed for the volume calculation are the integrated circuit thickness l_1, and the liquid-crystal variable waveplate thickness l_2.

The beam-steerer has angular resolution $\Delta\theta$ equal to the Bragg selectivity, which (in air) is given by $\Delta\theta = 2\lambda/L$ for 2nd-null separation, where λ is the wavelength in vacuum. The total angular swing allowed by the beam-steerer is θ in air. Because of Snell refraction, this angle is transformed into θ_g and θ_c inside the beam-splitter and crystal respectively (refractive indices n_g, n_c). Therefore, the number of holograms that can be stored inside the crystal in this architecture is

$$M = \frac{\theta_c}{\Delta\theta_c} = \frac{n_g \theta_g L}{2\lambda}. \tag{6.2}$$

The signal beam undergoes appreciable Fresnel diffraction because of the small pixel size. The diffraction spreads χ_g, χ_c (in glass and crystal, respectively) are determined by scalar diffraction theory according to

$$\sin\chi_g = \frac{\lambda}{n_g b} \qquad \sin\chi_c = \frac{\lambda}{n_c b}. \tag{6.3}$$

The sizes of the optical components can now be determined in terms of the DHR

and beam-steerer parameters as follows:

$$L_1 = \frac{L}{1-2\tan\theta_g}, \tag{6.4}$$

$$L_2 = \frac{1}{1-4\tan\theta_c\tan\theta_g}\left(\frac{2L\tan\chi_c}{1-2\tan\theta_g}+\frac{Nb}{\phi\,(1-2\tan\chi_g)}\right), \tag{6.5}$$

$$L_3 = \frac{1}{1-4\tan\theta_c\tan\theta_g}\left(\frac{L}{1-2\tan\theta_g}+\frac{2Nb\tan\theta_c}{\phi\,(1-2\tan\chi_g)}\right), \tag{6.6}$$

$$L_4 = \frac{Nb}{\phi\,(1-2\tan\chi_g)}. \tag{6.7}$$

Let $L_{m,x}$, $L_{m,y}$, $L_{m,z}$ denote the dimensions of the basic module. From Figure 6.8b we obtain

$$L_{m,x} = l_1 + L_1 + L_2, \tag{6.8}$$

$$L_{m,y} = \max\{L_2\ ,\ h\}, \tag{6.9}$$

$$L_{m,z} = \begin{cases} l_1 + L_3 + L_4, & \text{if } l_2 < \dfrac{L_3 - L_1}{2}, \\ l_1 + l_2 + \dfrac{L_1 + L_3}{2} + L_4, & \text{otherwise.} \end{cases} \tag{6.10}$$

The modules are arranged in grids as shown in Figure 6.8c, each grid sharing the same laser diode (or VCSEL) source. The grid contains G_x basic modules in the x direction and G_z in the z direction (see figure). The density and access time would be better in an arrangement with one source per basic module, because some excessive optical components (the peripheral beam-splitters delivering the beams to the modules in Fig. 6.8c) would not be required, and in addition it would be possible to read out each individual module simultaneously. However, source sharing alleviates problems due to excessive heat dissipation in the system, and also reduces the cost. Thus it seems that a small grid such as $G_x \times G_z = 2 \times 2$ is conservative enough for the heat and cost concerns without degrading the density and access time by much.

The grid volume is straightforward to calculate from Figure 6.8c. If we let w denote the size of the laser source, then

$$V_{grid} = (L + G_x L_{m,x})\, L_{m,y}\, (w + L + G_z L_{m,z}). \tag{6.11}$$

Parameter	Upper bound	value	Lower bound	value
N	complexity of the circuitry	2000		1
b	maximum die size L	1 cm	minimum feature size	4μm
L	maximum aperture, maximum DHR die size	1 cm	N/A	
θ	minimum feature size	10° (in air)		0°

Table 6.2: Constraint bounds for the volume optimization parameters.

The total number of (raw) bits that can be stored in the $G_x \times G_z$ grid is $G_x G_z M N^2$; therefore, the volume density of the memory is

$$\mathcal{D} = \frac{G_x G_z M N^2}{V_{\text{grid}}}. \tag{6.12}$$

This is our final result for the density. We seek to optimize $\mathcal{D}$ against the parameters N, b, θ, and L. Each one of them is constrained by technological limitations. A summary is given in Table 6.2.

Parameter	Value
λ	670nm
n_g	1.5
n_c	2.3
w	5mm
l_1	3mm
l_2	1.5mm
ϕ	0.5
$G_x \times G_z$	2×2

Table 6.3: Parameters used for the density and volume calculations.

Parameter	Value
Number of DHR pixels N	1,250
DHR pixel size b	4μm
Beam-steerer angular swing θ	10° (in air)
Laser beam diameter L	1cm
System density $\mathcal{D}$	36.0 Tbits m^{-3}
1 Tbit system volume	(30.3cm)^3
Number of grids and lasers N_g	123
Number of basic modules	492

Table 6.4: Results of constrained density optimization.

The remaining parameters were considered fixed and are given in Table 6.3. It is instructive to calculate the required volume of a Tbit system constructed with our proposed architecture. Figure 6.9 shows the result versus N and b, obtained by fixing L and θ to their maximum values of 1cm and 10° respectively, and ignoring (for the moment) the technological limitations on N and b. As we observe, the volume is minimized for the optimal combination $N = 3,278$, $b = 1.53\mu$m. The density at this

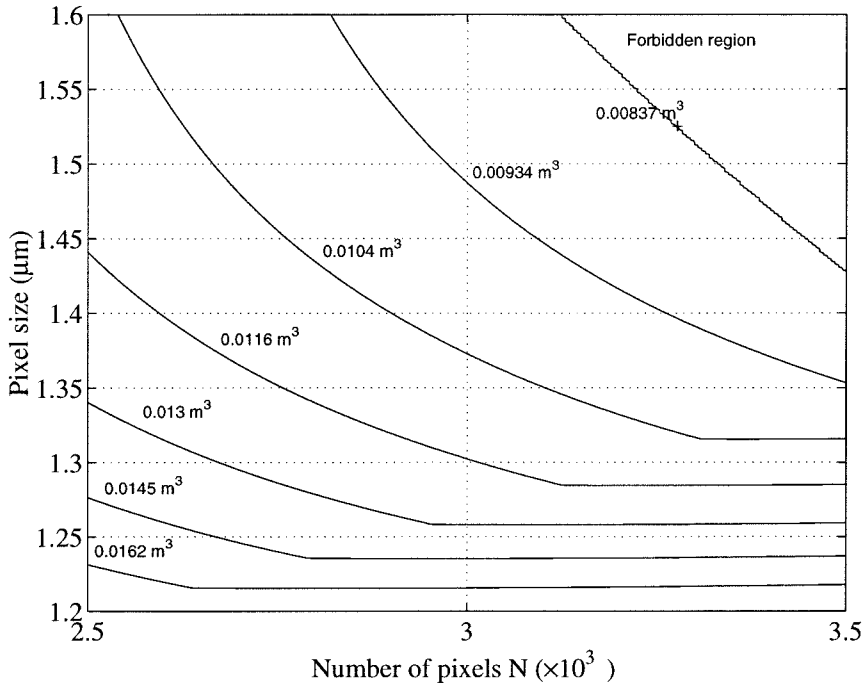


Figure 6.9: Volume of a Tbit modular memory constructed according to the architecture of Fig. 6.8 and the parameters of Table 6.3.

point is 119.5 Tbits m^{-3}, and the system occupies a volume equivalent to that of a cube of dimension 20.3cm. If b decreased from its optimal value, then the overhead due to the diffraction spread would overcome the gain in density because of the larger number of pixels that would fit within the DHR die, and hence the density would decrease. If, on the other hand, b increased to more than its optimal value, then $\mathcal{D}$ would again decrease, this time because N would decrease fast enough to overcome the reduced diffraction spread.

Unfortunately, the optimum found above is outside the feasible range of current technology. By solving the fully-constrained optimization problem, we find that it is best to fix b to its minimum value, θ to its maximum value, and set $N = \phi L/b$. The complete results are given in Table 6.4.

6.2.4 Noise, probability of error, and data rate

Here we consider the effects of optical and electrical noise on the compact holographic memory (see also section 2.2.1). The optical noise has μ effective degrees of freedom determined by the DHR pixel size b, the crystal half-aperture $R = L_4/2$, and the distance $d = L_4$ separating the aperture from the detector (see section 2.2.2). Substituting into (2.56) and (2.57), we obtain

$$\mu = \left(\frac{b}{4.88\lambda}\right)^2. \tag{6.13}$$

For the optimal $b = 4\mu$m calculated above, we have $\mu = 1.5$. The optical SNR is fixed by the properties of the optical system and the holographic material. For the data of Fig. 6.7d, using definition (2.30) we measured $(\mathrm{SNR})_{\mathrm{opt}} \approx 38.6$. We will consider this to be the upper limit on the optical noise performance.

The electrical SNR is determined by the noise behavior of the DHR circuit [136] and depends on the laser readout power P_{ref} and the detector integration time τ as

$$\left(\frac{(M/\#)^*\, L}{MN}\right)^2 P_{\mathrm{ref}}\,\tau = 1.236 \times 10^{-7} \times (\mathrm{SNR})_{\mathrm{el}}^2 \left(1 + \sqrt{1 + \frac{5208}{(\mathrm{SNR})_{\mathrm{el}}^2}}\right), \tag{6.14}$$

where P_{ref} is expressed in mWatts, and τ in μsec. Here $(M/\#)^*$ is a system parameter that we call "specific M-number." It expresses the $(M/\#)$ system metric [91, 110] for a holographic medium of unit-length. Since in the regime of weak holograms the diffraction efficiency increases with hologram thickness approximately as L^2 [138], the relation

$$(M/\#) = (M/\#)^*\, L \tag{6.15}$$

provides a convenient means of expressing the effect of material thickness on the dynamic range of an otherwise invariant optical system. Fixing the electrical SNR to $(\mathrm{SNR})_{\mathrm{el}} = 10$ (a realistic value for good noise performance), we can calculate the bit

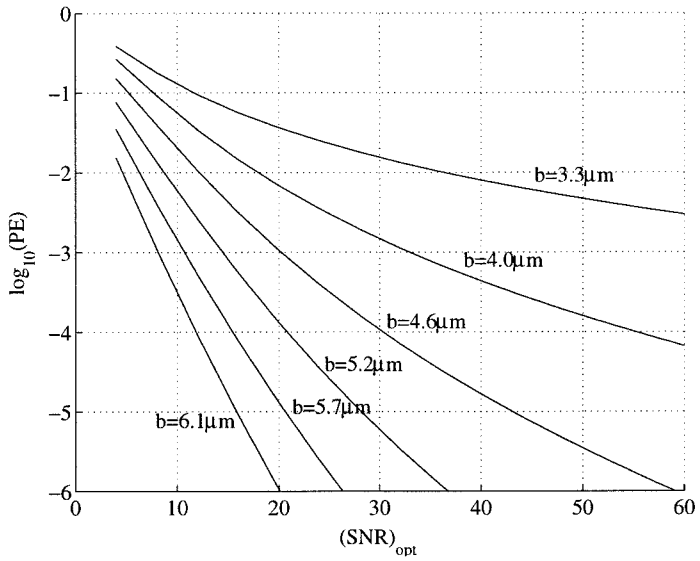


Figure 6.10: Probability of Error (PE) as function of signal to noise ratio (SNR) for different pixel sizes.

error rate versus $(\text{SNR})_{\text{opt}}$ and b according to the theory of sections 2.2.1 and 2.2.2. The result is given in Figure 6.10. Observing the numerical values we note that, for the value of b ($b = 4\mu\text{m} \Rightarrow \mu = 1.5$) obtained from the volume optimization, $(\text{SNR})_{\text{opt}}$ must be significantly high in order to achieve reasonably low PE. For example, if we set PE=10^{-4} as threshold, we would require $(\text{SNR})_{\text{opt}} \approx 55$. This is rather optimistic.

We can make the design more realistic by trading off some density for better error rate. From Figure 2.1 we observe that $\mu = 2$ ($b = 4.6\mu$m) allows PE=10^{-4} to be achieved with only $(\text{SNR})_{\text{opt}} \approx 30$.

We are now ready to calculate the data rate by considering that only one module per grid may be read out at one time, but that all grids may be read out in parallel. The result is

$$\mathcal{R} = 32.3 \times \left(\frac{(M/\#)^* \lambda}{n_g \theta_g (\text{SNR})_{\text{el}}} \right)^2 \frac{N_g P_{\text{ref}}}{1 + \sqrt{1 + \dfrac{5208}{(\text{SNR})_{\text{el}}^2}}} \quad \left(\frac{\text{Gbits}}{\text{sec}} \right), \tag{6.16}$$

where N_g is the number of grids in the entire system, and the units are mm^{-1} for $(M/\#)^*$, μm for λ and mWatts for P_{ref}. Collecting all the results of the previous

Parameter	Value	Parameter	Value
N	1,081	$\mathcal{D}$	27.9 Tbits m^{-3}
b	4.6μm	1 Tbit system volume	$(33.0cm)^3$
θ	10° (in air)	N_g	164
L	1cm	$\mathcal{R}$	708.1 Gbits/sec

Table 6.5: Final design of 1 Tbit compact holographic memory.

optimizations, we obtain the final design summarized in Table 6.5.

Note that the above calculations refer to the *instantaneous* rather than the sustained data rate. They do not take into account the settling times for the liquid crystal devices, and the speed of the data bus connected to the DHR's of the modules.

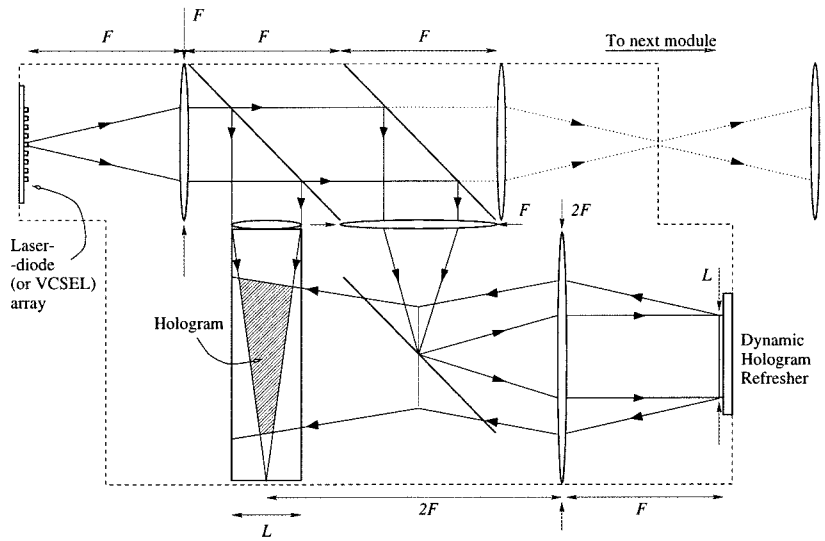
6.2.5 Shift-multiplexed compact module

As mentioned already, beam-steerers are expensive and difficult to fabricate. In addition, they have other practical problems, e.g., they are slow (because they are limited by the response time of the liquid crystals), and they diffract at all orders thus introducing significant first-order crosstalk between holograms. A very good alternative to beam steerers is laser (VCSEL or diode) arrays. Then instead of steering a single beam to access different holograms, we associate one source with each hologram and use them one at a time. Apart from solving the beam steering problem, the laser array opens some interesting possibilities: e.g., we can do simultaneous angle and wavelength multiplexing in the compact architecture, and increase the sustained data rate of the system, since the lasers can be switched very fast (at GHz rate). One problem with using multiple sources is producing the signal beam for recording[5], which of course must originate at the same source that produces the reference[6]. Figure 6.11 gives an example of a compact architecture with a laser array that employs shift multiplexing.

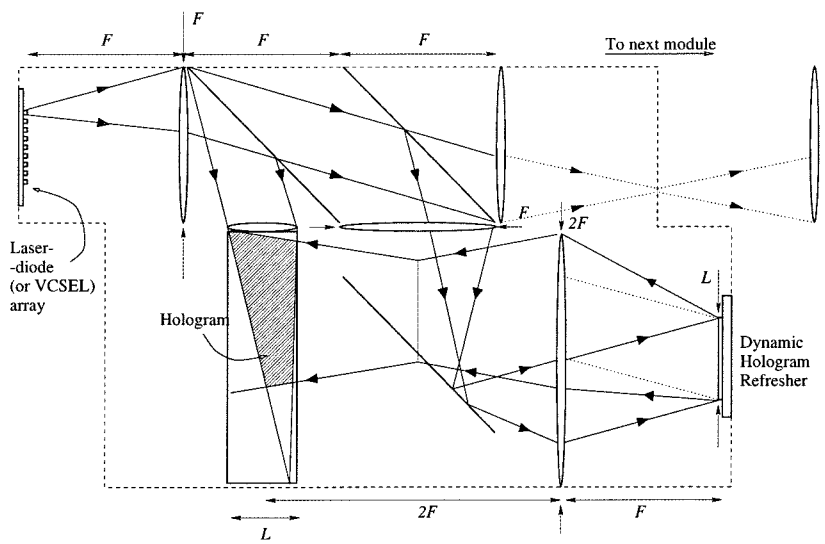
Let d denote the spatial separation between adjacent laser sources in the array,

[5]Read-only architectures with laser arrays are very compact and easy to design, but we will not consider them in this thesis.

[6]Two different phase-locked sources may be used; however, phase-locking requires additional bulky optical elements, e.g., isolators.



(a)



(b)

Figure 6.11: Compact holographic memory design utilizing a laser source array and shift multiplexing. (a) Beam paths for the hologram corresponding to the central laser source; (b) Beam paths for the edge source.

and a the aperture of the sources. Using the theory of section 3.3.1, we find that the shift selectivity in the arrangement is $\delta = 4\lambda F/L$. It is straightforward then to derive the number of sources M that can be accommodated by the optical system, and therefore the number M of stored holograms:

$$M = \frac{L}{2}\left(\frac{1}{\lambda} - \frac{1}{a}\right) \tag{6.17}$$

If we assume that the sources can be packed as densely as allowed by $a = d/2$, then we obtain that the capacity is maximized when $F = 2L$ to the value $M_{\max} = L/4\lambda$. The volume density of this system is given by:

$$\mathcal{D} = \frac{L^3\phi^2}{24F^3b^2}\left(\frac{1}{\lambda} - \frac{1}{a}\right). \tag{6.18}$$

Using the optimal F derived above, $b = 2\mu$m, and $\phi = 0.5$ we find that $\mathcal{D} = 60.73$ Tbits/m^3 can be achieved by this system with $M = M_{\max} = 3,731$ holograms. Such a system would have surface density ≈ 233 times higher than a silicon DRAM with the same pixel size (2μm). The large number of holograms is, however, a concern for two reasons: (a) it would be challenging to fabricate such a large array of vcsels; (b) the dynamic range of the holograms would be very bad unless a high-$(M/\#)$ material were available. It is more reasonable to reduce the number of holograms (and the density) by a factor of, say, 10, obtaining

$$M = 373, \qquad \mathcal{D} = 6.07 \text{ Tbits/m}^3.$$

This design offers surface density higher than silicon DRAM by a factor of 23. The required system volume for 1 Tbit is a cube of side 54.8 cm.

6.2.6 Discussion

The cost of the basic dynamic memory module has been calculated and optimized versus current and projected industry standards in very similar architectures [136]. For

the angle-multiplexed architecture, we estimate roughly \$170 per module; therefore, the cost of the optimized Tbit system[7] would be \$518,000, or \$4.14/MByte. It turns out that the beam-steerer is a major component (approximately 50%) of the cost. We do hope that the reflective beam-steerer design will reduce this cost, thus making the entire system cheaper. It is harder to estimate the cost of the shift-multiplexed architecture, because the vcsel with the requirements we posed are not commercially available yet.

Several improvements may be made on the designs we presented in sections 6.2.3 and 6.2.5. For example, rather than the grid architecture of Figure 6.8c, one may arrange a linear architecture, where each laser is feeding a row of basic memory modules. This saves space because the peripheral beam splitters of Figure 6.8c that deliver the reference and signal beams to the modules would not be needed. A calculation along the lines of section 6.2.3 shows that the achievable density with the same parameters of Table 6.3 is 38.9 Tbits m^{-3}, in other words a memory of 1 Tbit would fit in a cube of dimension 29.5 cm. However, the linear architecture does not have the nice property of equalized path lengths that we observe in Figure 6.8c. Therefore, a bulkier and/or more expensive source might be required for higher coherence, which would offset the density improvement.

Further improvement may be obtained by inserting a lens in the signal beam path, between DHR and PBS2, or between PBS2 and the storage crystal, as shown in Figure 6.12. This makes the optical system similar to the one we analyzed in section 5.1.2, proposed by van der Lugt [103] in 1975 for optimizing holographic storage. The lens may be of very poor quality, since the phase-conjugation process will undo all the aberrations it might induce in the signal beam. Therefore, the volume taken by this lens, and the cost it adds to the system, would be immaterial. On the other hand, big gains are made in density and cost (since the volume of the photorefractive crystal would decrease) because the lens would reduce the signal beam spread.

[7]The cost calculation does not include interfaces, packaging, or marketing.

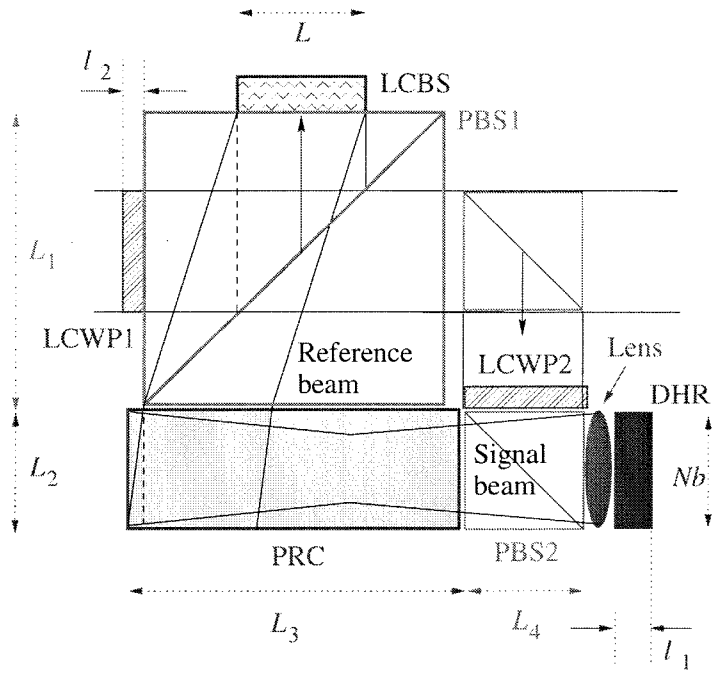


Figure 6.12: Compact holographic memory utilizing van der Lugt's idea for reducing the signal beam spread.

Referring to Fig. 6.12, we assume that the following conditions hold:

$$F = \frac{Nb^2}{2\lambda} \qquad L_3 + L_4 \leq 2F.$$

This assures us that the signal beam never defocuses beyond L until it reaches the edge of the hologram, and that we can use $L_2 = L_4 = L$. The remaining two dimensions are calculated as

$$L_1 = \frac{L}{1 - 2\tan\theta_g}, \tag{6.19}$$

$$L_3 = L\left(\frac{1}{1 - 2\tan\theta_g} + 2\tan\theta_c\right). \tag{6.20}$$

Repeating the volume optimization we find that this system offers maximal (under the constraints of section 6.2.4) density $\mathcal{D} = 34.2$ Tbits/m^3 if the modules are arranged in grids, and $\mathcal{D} = 49.4$ Tbits/m^3 if they are arranged in rows. The volume of a Tbit

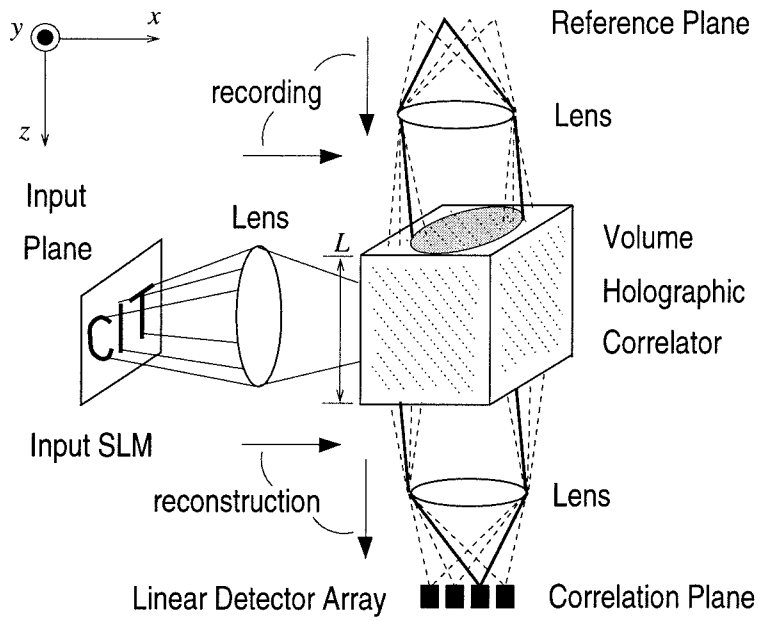


Figure 6.13: Volume holographic correlator in the 90° geometry.

system is $(30.8\text{ cm})^3$ and $(27.2\text{ cm})^3$, respectively.

Finally, one should mention that the analysis presented in section 6.2.3 contains several implicit approximations. For example, we did not take into account the effects on selectivity and diffraction efficiency of the displacement introduced to the phase-conjugate reference by the mirror coating. Several issues about the approximations will become clearer as the practical implementations progress further.

6.3 Associative memory access

6.3.1 Van-der-Lugt correlators

Consider[8] the 90° geometry layout of Figure 6.13. Each hologram is recorded using the SLM to display the signal beam, and one of the reference beams generated in the reference plane, e.g., by a rotating mirror and a lens (not shown). Let $S_n = \tilde{f}_n$ denote the Fourier transform of the n-th stored image, and $R_n = \mathrm{e}^{ik_n x}$ the n-th reference.

[8]I thank Michael Levene for many illuminating discussions on the topic of shift invariance in volume holographic correlators.

The n-th hologram is then expressed as the inteference pattern

$$|R_n + S_n|^2 = 1 + \left|\tilde{f}_n\right|^2 + \mathrm{e}^{ik_n x}\tilde{f}_n^* + \mathrm{e}^{-ik_n x}\tilde{f}_n \tag{6.21}$$

During the readout phase, rather than illuminating the hologram with one of the reference beams, we use the SLM to illuminate the hologram with an image g. Then the 3rd term in (6.21) is Bragg-matched for all holograms[9] ($n = 1, \ldots, N$), yielding on the detector plane (x', y') the reconstruction

$$c(x', y') = \sum_n \mathcal{R}_{f_n g}\left(x' - \frac{\lambda F k_n}{2\pi}, y'\right), \tag{6.22}$$

where $\mathcal{R}_{fg}$ is the cross-correlation of two functions, defined as

$$\mathcal{R}_{fg}(x', y') = \iint_{-\infty}^{+\infty} f(z, y) g^*(z - x', y - y')\mathrm{d}z\mathrm{d}y. \tag{6.23}$$

Therefore, the holographic memory in this mode acts as a matched filter that compares the input with all stored templates simultaneously. Since (6.22) is a linear relationship, we can think of the holographic correlator as interconnecting the input (SLM) plane with the output (correlation) plane via weights determined by the holograms. Typically, the detector outputs are passed to a winner-take-all circuit which determines the pattern f_{n_0} that matches the input g most closely. Therefore, the system of Figure 6.13 is a general-purpose pattern recognition device. Variants of this system have been used in a variety of applications, e.g., face recognition, target detection, fingerprint recognition, cryptography, and robot navigation.

Important for the characterization of a pattern recognition system are its *invariances*, i.e., the set of transformations that can be applied to the input without altering the result of the matching operation. Neglecting volume diffraction effects for the moment, the volume holographic correlator of Figure 6.13 is *shift*-invariant in the y direction. Indeed, assume that the input is $g(x, y) = f_{n_0}(x, y - \alpha_y)$. Then from

[9]The 1st, 2nd, and 4th terms are, to good approximation, Bragg-mismatched in a volume hologram.

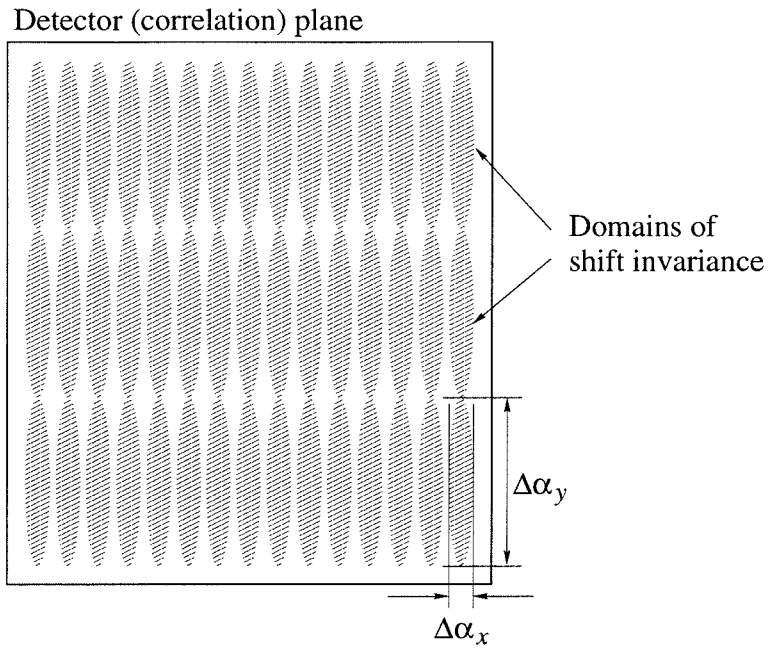


Figure 6.14: Correlator grid in the 90° geometry.

(6.22) we deduce that the output is simply $\mathcal{R}_{f_{n_0}}(x', y' - \alpha_y)$, which is maximized at $(x', y') = (0, \alpha_y)$. On the other hand, the shift invariance in the x direction is limited by the spacing of the correlators, because if the input is shifted by a large amount, then it would enter the domain of a different filter thus yielding the wrong result.

Volume diffraction modifies the properties of the optical correlator because of the effect of Bragg mismatch. The simplest way to understand the effect of Bragg mismatch on shift invariance is to reverse the rôles of signal and reference beams. If the input image is shifted by an amount (α_x, α_y) at the SLM plane, then its Fourier transform is angularly detuned by $(\Delta\theta_x, \Delta\theta_y) = (\alpha_x/(\lambda F), \alpha_y/(\lambda F))$, where F is the focal length of the lenses in the reference beam path. The correlation results from reconstruction of the hologram by the signal beam; therefore, Bragg mismatch occurs when the signal is shifted. The shift invariance is equivalent to the angular Bragg selectivity of the reference beam, and is asymmetric in the x, y directions:

$$\Delta\alpha_x = \frac{\lambda F}{2\mathrm{L}} \qquad \Delta\alpha_y = \sqrt{\frac{2\lambda}{L}}\, F \tag{6.24}$$

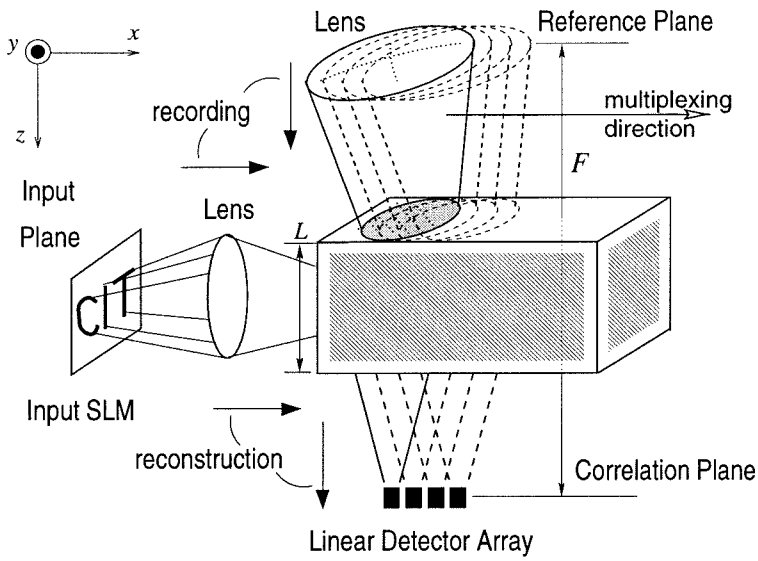


Figure 6.15: Shift-multiplexed volume holographic correlator.

Thus the shift invariance domains of an angle-multiplexed 90° correlator are arranged like a rectangular grid of elipses, as shown in Figure 6.14. Notice the asymmetry in the two directions, due to the different Bragg selectivities, and also the length L that enters the calculation and is measured in the direction perpendicular to the one used for angular selectivity in memories (measured along the signal axis).

6.3.2 Shift-multiplexed holographic correlators

The idea of shift multiplexing is applied to correlator design simply by replacing the plane wave reference beam with a spherical beam, as shown in Figure 6.15. During recording, the reference beam is translated relative to the recording medium by a fixed amount δ_x in the x direction, as in the case of a shift-multiplexed memory. If we now read out the hologram with an input pattern g projected in the SLM, then we obtain the correlation of g with all the stored templates f_n simultaneously at the detector plane, in similar fashion as in the angle-multiplexed case, with one major difference: the correlation is formed without the need for an additional Fourier transforming lens (compare Figures 6.13 and 6.15).

It is straightforward to prove, using the theory of section 3.3.2, that the shift in-

variances in the system of Figure 6.15 are also[10] given by (6.24). If we perform shift multiplexing along both the x and y directions, then we obtain an asymmetric rectangular correlator grid identical to that of Figure 6.14. Therefore, shift multiplexing offers an interesting method of building optical correlators saving one lens compared to the "usual" angle-multiplexed design. Experimental shift-multiplexed correlator systems have not been built yet.

6.3.3 Compact architectures

We now combine the ideas of holographic correlators (section 6.3.1) and compact memories (section 6.2) to design an architecture that fits in a reasonable volume and combines the capabilities of recording, refreshing, and recalling holograms both by address (as a memory) and by content (as a correlator combined with a winner-take-all function). The architecture is sketched out in Figure 6.16.

Recording is performed using beams R and S. S is spatially modulated by DHR #1. Because of the presence of lens #1, the hologram is Fourier-transform type. The beam steerer is used to select the angle of incidence of the reference beam, so angular multiplexing can be performed. The geometry is chosen such that for every distinct angle of incidence of the reference, the reference beam R comes to a focus on a different pixel of DHR #2 after passing through lens #2. To reconstruct a particular hologram, the beam steerer is used to select the appropriate angle, and the corresponding pixel of DHR #2 is turned on. As a result, the conjugated reference beam R^* (counter-propagating with respect to R) is generated and illuminates the hologram giving rise to the phase-conjugated reconstruction S^* which counterpropagates and comes to a focus on DHR #1, where it is detected according to the description given earlier. Refreshing is performed by detecting a page from the reconstruction of beam S^*, generating a stronger signal S and re-recording a hologram with S and R. Associative recall requires two steps: first the pattern g to be associated is generated on DHR #1 and illuminates the hologram. It is well known that, in this case, the

[10]Note, however, that the quantity F has a different interpretation.

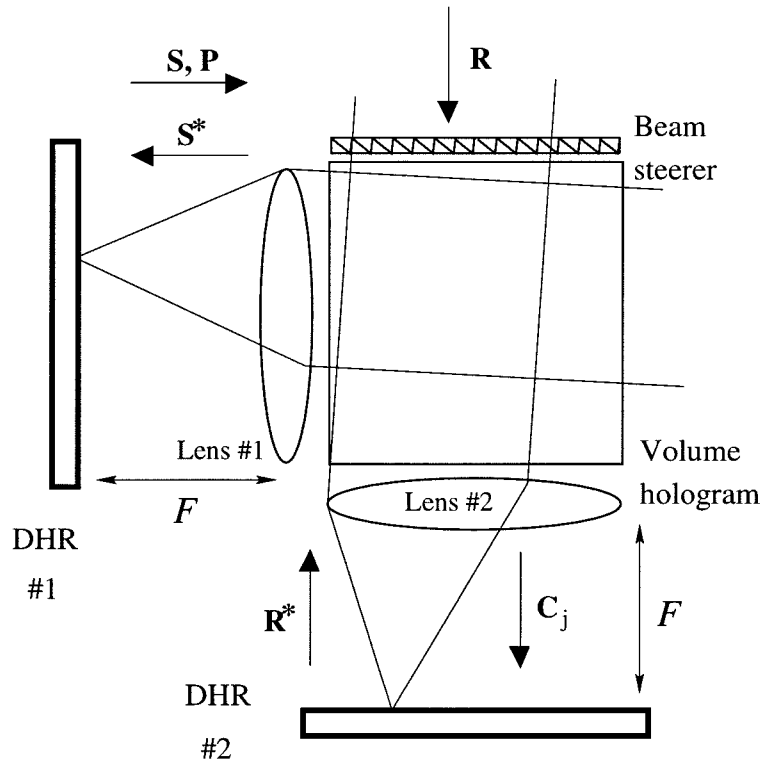


Figure 6.16: Compact architecture for a holographic memory with recording, refreshing, direct and associative recall capabilities.

correlations $\mathcal{R}_j$ between g and the stored images f_j will form simultaneously on the back focal plane of lens #2. According to the design restriction stated earlier, each correlation peak forms on the corresponding pixel of DHR #2, where the intensity is detected. Then a winner-take-all operation is performed to determine which stored pattern matches g best. This completes the first step. In the second step, the winning pixel of DHR #2 is turned on and the beam steerer is configured appropriately to generate a conjugated reference R^* that will read out the best match to pattern g, and will form the conjugated reconstruction S^* back onto DHR #1 where it is detected, and constitutes the result of the associative recall operation. Additional features, e.g., thresholding, multiple associations, etc., are easily added by programming the DHR's properly, or with external circuitry.

The presence of two lenses makes the architecture a bit less compact than that of the pure memory; however, the additional capability of associative recall is very useful

in high-performance computing applications. Many variations of this architecture using, e.g., different multiplexing techniques, removing one or both of the lenses for reduced size at the expense of density (if lens #1 is removed) or shift invariance (if lens #2 is removed), etc., are possible. An exhaustive analysis is beyond the scope of this work.

6.3.4 Space and time-domain correlators

So far we have described holographic memories with at most three degrees of freedom: "in-plane" multiplexing, "out-of-plane" (fractal) multiplexing, and spatial multiplexing (or using multiple modules as in section 6.2). An exciting possibility for enhancing the capacity of holographic memories is to introduce time-domain holography. We will describe one simple case, the three-photon echo. Many extentions and alternatives are possible [139, 140, 141, 142, 143].

Suppose[11] that two optical pulses with relative time delay τ illuminate a photosensitive medium. The pulse amplitudes are described by

$$a_R(\boldsymbol{r}, t) = v_R(\boldsymbol{r})u(t), \tag{6.25}$$

$$a_S(\boldsymbol{r}, t) = v_S(\boldsymbol{r})u(t-\tau), \tag{6.26}$$

where $v_R(\boldsymbol{r})$ and $v_S(\boldsymbol{r})$ are the spatial variations of the two pulses, and $u(t)$ is a simple pulse shape, e.g., a Gaussian or a hyperbolic secant (sech). In a particular location $\boldsymbol{r}_0$, the electric field is described as

$$a(\boldsymbol{r}_0, t) = [v_R(\boldsymbol{r}_0) + v_S(\boldsymbol{r}_0)]\ [u(t) + u(t-\tau)] \xrightarrow{\mathcal{F}_t} [v_R(\boldsymbol{r}_0) + v_S(\boldsymbol{r}_0)]\ U(\omega)\ \left(1 + \mathrm{e}^{-i\omega\tau}\right),$$

where $\mathcal{F}_t$ denotes Fourier transformation in the time variable. Suppose also that the medium absorbs over a broad band of optical frequencies, and the absorption is proportional to optical intensity. This is the case in spectral hole-burning materi-

[11] I thank Tom Mossberg, Alan Johnson, Roger McFarlane, and Christophe Moser for help on the contents of this section.

als [140, 144]. Assume that $u(t)$ is ultra-short so that the pulse spectrum is approximately flat over the entire absorption bandwidth of the optical medium. Then after exposure to the two pulses, the frequency spectrum of the absorption at $\boldsymbol{r}_0$ will have the form

$$\alpha(\boldsymbol{r}_0) = \alpha_0 + \alpha_1 \cos(\omega\tau), \tag{6.27}$$

where all the constants were lumped into α_0 and α_1. This looks like a grating in the frequency domain, and indeed it acts as one: suppose that we illuminate the exposed material with a third pulse $u(t-T)$, which also has a flat spectrum. As the pulse propagates in the medium, some of its frequencies are selectively suppressed according to (6.27). In the thin film approximation, the spectrum $V(\omega)$ at the output equals the product of the pulse spectrum and the absorption spectrum of the optical material, resulting in

$$Q(\omega) = [\alpha_0 + \alpha_1 \cos(\omega\tau)]\, \mathrm{e}^{-i\omega T} \propto \alpha_0 \mathrm{e}^{-i\omega T} + \alpha_1 \left(\mathrm{e}^{i\omega(t-T-\tau)} + \mathrm{e}^{-i\omega(t-T+\tau)}\right). \tag{6.28}$$

Taking the Hilbert transform (so that causality is obeyed) we obtain the output

$$q(t) = \alpha_0 u(t-T) + \alpha_1 u(t-T-\tau). \tag{6.29}$$

Therefore the frequency grating recorded by the two pulses interacted with the third pulse to generate a fourth pulse, a replica of the second pulse. In other words, we obtained in the time domain the equivalent of hologram reconstruction!

Two important restrictions on the time scales must be mentioned. The separation τ between the first pulses must be short enough so that the fringes of (6.27) are not finer than the frequency selectivity of the medium, i.e., the homogeneous bandwidth of the absorbing particles. On the other hand, T must not be longer than the dephasing time of the medium, because if the coherence between different frequency components of the time-domain hologram is destroyed, then the scattered light cannot take the

form of a delayed pulse, but would rather look random, like speckle in the time domain.

Information can be stored in time-domain holograms in one or both of two ways: (a) in the time domain by using a pulsetrain rather than a single pulse as signal a_S; (b) in the space domain by using, e.g., $v_R(\boldsymbol{r}) = \mathrm{e}^{ik_R x}$ and $v_S(\boldsymbol{r}) = f(\boldsymbol{r})\mathrm{e}^{ik_S x}$, when two gratings are recorded simultaneously: one in the time domain as described immediately before, and one in the space domain according to section 3.1. All properties of volume holograms, e.g., Bragg selectivity, carry over to time-domain holograms. Thus it is possible, for example, to angle-multiplex several pulsetrains, so that each pulse contains a different page of data. In this memory, time takes the place of a fourth dimension for storage. A pulsetrain may be correlated simultaneously with all stored waveforms in time and space and the correlations obtained simultaneously. Without getting into details, the possibility of performing associative recall in the time domain was first mentioned by Longuet-Higgins and Gabor in a series of articles in 1968 [145, 146, 147]. Several interesting possibilities exist of storing time sequences of events and recovering them from partial realizations.

Before concluding this section, we must mention some challenges in time-domain memories. Currently it seems that the spectral hole-burning property occurs only in temperatures of a few Kelvin; therefore, the complication of a cryostat is unavoidable for the experiments. The storage lifetimes of most spectral hole-burning materials are only in the order of a few milliseconds. Ultra-fast spatial light modulators do not currently exist that could modulate pulsetrains of M$\sim$GHz repetition rates. Similarly, ultra-fast detection is usually performed with autocorrelators, which have poor spatial resolution. Still, research in time-domain holographic memories is currently very active because of their fascinating and elegant properties.

Chapter 7 Awareness-based computation

In this chapter[1], we are interested in designing intelligent systems that can monitor and interact with complex, variable, and poorly modeled environments. This task remains a challenge, particularly for systems that need to be controlled in real-time, such as autonomous robots, automated buildings, and traffic control in metropolitan areas. We describe an approach that sacrifices the time-consuming (and, for many physical systems, ill-defined) goal of searching for the global optimum in favor of a locally optimal solution in a small, restricted subset of the system space. This "region of interest" is determined in real-time as the best representation of the system status given limited computational resources, and changes as the system and the environment evolve. The organization of our model is reminiscent of the cognitive architecture of the primate brain, and makes use of a notion similar to the hypothesized function of awareness [148, 149].

We experimented with computation in unknown environments of high complexity using a computer game of *Desert Survival* as testbed. As we describe in section 7.1, the computer game is too complex to admit a solution (i.e., a consistently winning strategy) with traditional optimization techniques [150, 151, 152] in reasonable time. Instead, we produced interesting behavior by using a selective attention technique to dynamically identify important regions of the input space. The computational resources were subsequently concentrated at any given time in the currently selected region. The real-time contraint imposed an optimum on the size of the region of interest in terms of algorithm performance. The generalization of the model to wide-purpose computational problems is given in section 7.2, along with a comparison with

[1]I am grateful to R. A. Andersen, A. Batista, F. Crick, R. Denkewalter, P. Perona, and S. Shimojo, for numerous helpful discussions and suggestions on the contents of this chapter.

the primate cognitive system. The learning aspects of awareness-based computation, the connection to holographic memories, and a simplified theoretical analysis of the *Desert Survival* learning model are discussed next in section 7.3. A short discussion on future directions in the design of intelligent systems is given in the concluding section 7.4.

7.1 The *Desert Survival* simulation

7.1.1 Description of the computer game

We are seeking design principles for intelligent systems that can interact with highly complex and dynamic environments in the presence of high-dimensional inputs and massive memory. A common theme in these applications is that, at any given moment, numerous sub-functions, seemingly un-related to each other, are running in parallel to cope with a specific aspect of the problem at hand; yet, when viewed in its entirety, the system should display unified behavior, emerging from the cooperative interactions of the sub-functions or agents [153].

We generated an artificial complex environment in the form of a computer game of *Desert Survival*[2] taking place in a virtual desert, described in Figure 7.1. The players are two Sheiks, who fight to conquer as many oases as possible, and become richer as a result. Each Sheik has an army of camels at his disposal. Each camel is capable of navigating in the desert towards some oasis, managing its water (if a camel runs out of water, it dies) and trading water for money (each camel starts out with some money and water). An oasis switches its allegiance to one particular Sheik if the number of camels belonging to that Sheik in that oasis at this point in time exceeds the camels of the opposing Sheik by a fixed amount $T = 1$. If left alone, the camels are not capable of organizing their behavior as a team; the strategy that coordinates the actions of the camels is provided by the Sheiks.

[2]The *Desert Survival* simulation was programmed in C for X Windows version 11 release 6 under the Solaris 2.4 environment, and ran on a SparcStation 2. Movies taken from sample simulation runs are available at `http://sunoptics.caltech.edu/~george/DesertSurvival`. Robert Denkewater and Petru-Nicolae Chebeleu contributed to the graphics and animation.

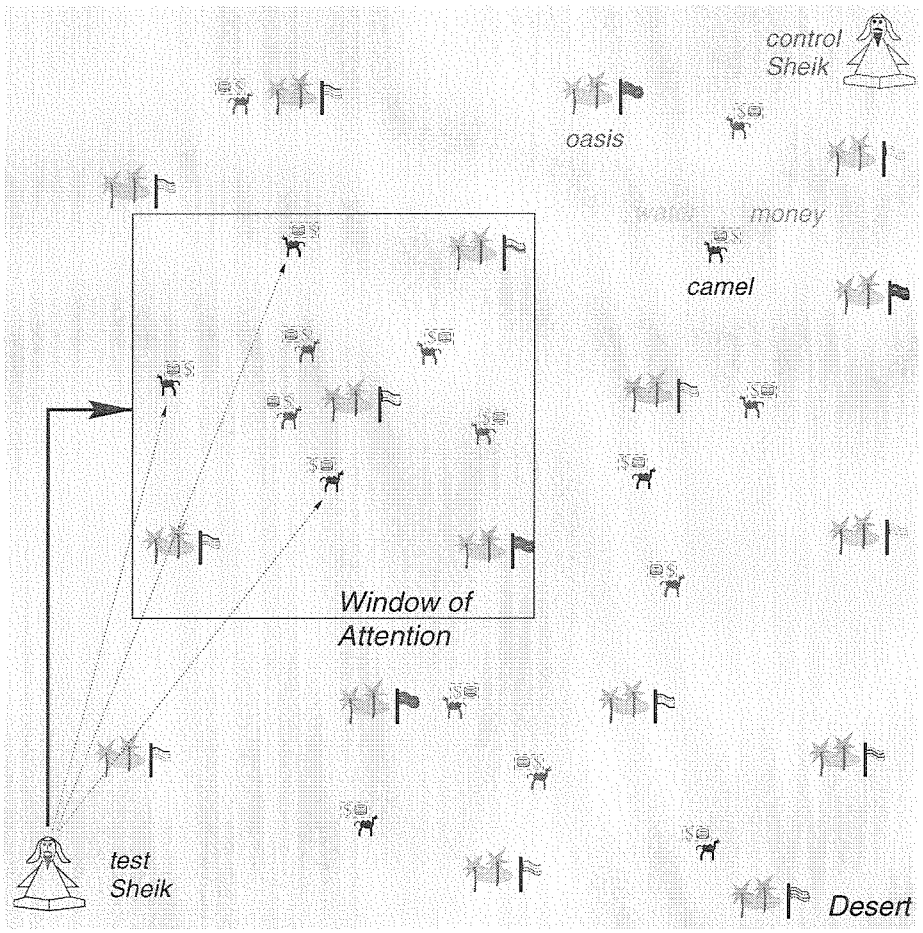


Figure 7.1: The *Desert Survival* simulation.

The implementation of *Desert Survival* presented in this chapter is played on a 100×100 grid, with 288 regularly-arranged oases, initially divided equally between the two Sheiks. Each oasis starts with $2,000 in cash and accumulates more money when it sells water to opponent's camels. Time advances in steps of two virtual hours, with the two Sheiks alternating their moves as in chess (the Sheik to move first is determined at random). At the beginning of the simulation, each Sheik has 50 camels, and each camel is given 100 gallons of water and $200. The camels are free to spend water and money as they please (the Sheik does not interfere with these decisions). The camels do not move continuously, but hesitate at each step: they advance initially with probability 0.1 and stay otherwise. Every time the camels

reach an oasis, their moving probability is increased by 5%, simulating procedural learning (that is, improving agent performance through individual experience). Once a camel is within an oasis, it likes to stay inside, and moves out with probability 0.2. As the camels move, they drink water from their supplies at a rate that depends on the time of day and their activity (the rate varies from 0.1 gallon per hour when they rest at midnight to 1 gallon per hour when they walk during the hot afternoon hours). They refill their supplies with water that they find at oases, but have to pay for it ($10 per gallon) if the oasis belongs to the opponent Sheik. If a camel runs out of water in the desert, it dies. When both Sheiks are equivalent, the death rate is 0.05 camels per day. It can be much higher, though, if one of the Sheiks is at a serious disadvantage compared to his opponent.

The complexity of this game seems at first overwhelming, because of the numbers of agents (see the previous paragraph) and the long duration (typically several virtual weeks). To solve *Desert Survival*, we developed a new computational method that could lead to substantial improvements over existing techniques, in poorly modeled, non-stationary environments, where overwhelming rates of information are received from sensors and memory, as it often happens in the real world. Rather than looking at the entire desert, the Sheiks use a reduced amount of information which they extract and process in a two-step procedure.

The first step involves the formation of a so-called saliency map [154], computed on the oasis locations. The saliency of an oasis is a real number defined as

$$s_k = \begin{cases} \tilde{C} - C - T & \text{if the oasis belongs to the Sheik,} \\ \tilde{C} - C + T & \text{if the oasis belongs to the Sheik's opponent,} \end{cases} \tag{7.1}$$

where C is the number of Sheik's camels in a 3×3 window around the oasis, and $\tilde{C}$ is the number of the opponent's camels within the same neighbourhood.

From definition (7.1), it follows that the saliency s_k of oasis k expresses the likelihood that the oasis will be overtaken within a few (typically two to three) time steps. The saliency values s_k are computed according to (7.1), and are subsequently aver-

aged over an $A \times A$ moving window. The most salient $A \times A$ region is then selected, and an $A \times A$ "window of attention" is formed at the same location. Typically, A is far smaller than 100, as we will discuss later. Oasis ownership changes as camels move in and out of them, thus affecting saliency. The window of attention moves freely in response, keeping up with the changes.

In the second step, the Sheiks concentrate their efforts to the most salient region selected by attention. For each of his camels i, the Sheik computes a set of navigation instructions, expressed as the probabilities p_{ik} that the camel i will visit oasis k (located within the window of attention). The probabilities p_{ik} are computed from the heuristic formula

$$p_{ik} = \frac{1}{p_0} \times \frac{e^{\alpha s_k}}{1 + \beta d_{ik}}, \tag{7.2}$$

where d_{ik} is the Euclidean distance between the current position of the camel and the oasis, and p_0 ensures that $\sum_k p_{ik} = 1$. Each camel within the window then selects an oasis to head for, according to the probabilities p_{ik} that it was assigned by the Sheik. The parameter α controls territorial aggression by giving preference to salient locations; β determines how reluctant the camels are to undertake remote trips where they may perish from lack of water. The simulation experiments are divided into two classes: experiments without memory (section 7.1.2), where α and β are constants, optimized experimentally; and experiments with episodic learning (section 7.1.3), where the Sheiks are allowed to dynamically learn by experience the appropriate values of α, β in response to their inputs.

An additional requirement for the Sheiks is that they must complete all processing in limited time (like speed chess, and many other situations in the real world). The time limitation is implemented as follows: at the beginning of each time step, each Sheik is given 125 "tokens." The test Sheik spends $5A$ tokens to process the contents of his attention window, and in addition 5 tokens to update each camel (calculate the set of p_{ik}'s and transmit them). If during the calculations the Sheik runs out of tokens, he must abandon all processing; therefore, he is forced to "neglect" some of

his camels within the window of attention. These neglected camels are not assigned navigation instructions p_{ik}; they are either left to wander the desert at random, or continue to follow previously given but now outdated instructions.

What is the optimal size A of the attentional window? As indicated by the choice of name, its usage is reminiscent of the action of selective visual attention in focussing resources to a restricted portion of the visual scene [155, 156, 157]. The window size A controls how much information (the positions of the oases and camels[3]) is available to the Sheik in determining his strategy. Thus the Sheik is only "aware" of the values of the saliences s_k and the distances d_{ik} within the window of attention. Two extreme cases are obviously bad: if A is too small, the attentional window does not contain enough information and the algorithm is not effective. If, on the other hand, A is too large, the algorithm can work well (since more and more information is available to the Sheiks) but not sufficient time is available to evaluate this information.

7.1.2 Experiments without learning

To study the effect of A in the *Desert Survival* computer game, we disable one of the two opponents, the "control" Sheik (that is, we let his camels roam around unguided), and apply the probabilistic navigation algorithm described above with different A values to the second "test" Sheik only. We evaluate performance with the metric P, defined as the relative cash advantage of the test Sheik compared to the control Sheik:

$$P = 100 \times \frac{\mathrm{m}_{\mathrm{test}} - \mathrm{m}_{\mathrm{control}}}{\mathrm{m}_{\mathrm{test}} + \mathrm{m}_{\mathrm{control}}} \quad (\%)\,, \tag{7.3}$$

where $\mathrm{m}_{\mathrm{test}}$, $\mathrm{m}_{\mathrm{control}}$ are the total cash quantities available to the test or control Sheiks (summed over the camels and oases they own), respectively, after a fixed duration t of simulation time.

In this section we describe the first experiment, where no learning occurs; i.e., all Sheik functions are pre-programmed as described in the previous section, and no

[3] In principle, other quantities, such as the amounts of money and water, could be used as well.

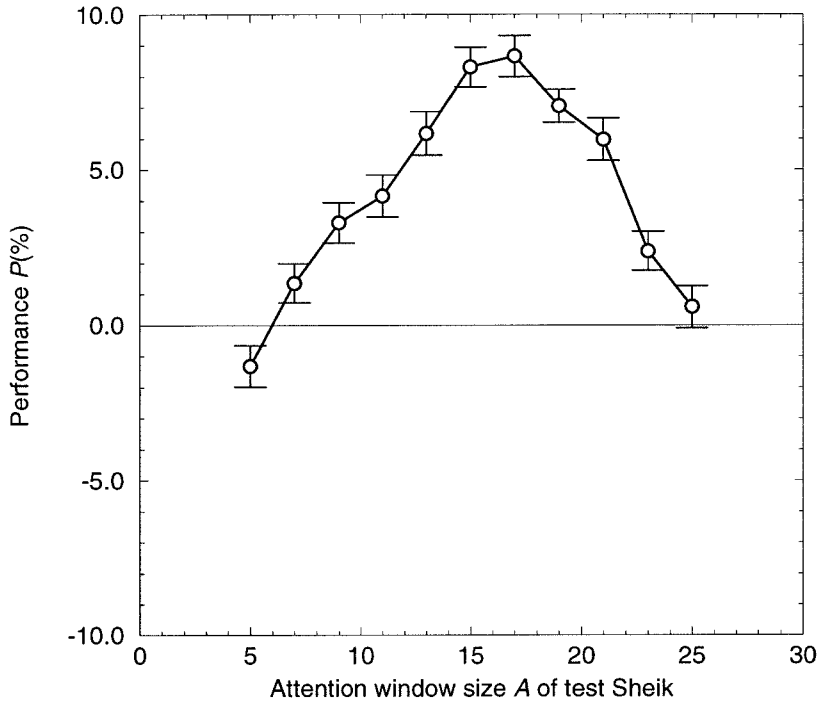


Figure 7.2: Performance P of the test Sheik without any learning ($\alpha = 0.1$ and $\beta = 1.0$) whose attention window size A is given by the abscissa, against the control Sheik, whose camels wander randomly. The error bars correspond to the standard deviation of P in 100 independent trials.

adaptation is allowed. The game duration is $t = 35$ days (=420 two-hour steps). The data for each A are derived from 100 independent trials of the same duration t. The results are shown in Figure 7.2. As expected, P is at or below chance for small A, peaks at around $15 \sim 17$, and drops to chance again for larger A. The peak shifts to the right if the test Sheik carries out the computations associated with (7.2) faster.

This experiment demonstrates the main point of restricting computation to a prudently selected "region of interest" when the computational resources are limited. The mechanism is reminiscent of the information reduction occuring in the primate brain thourgh the effect of attention, as we discuss in more detail in section 7.2.

7.1.3 Experiments with Episodic Memory

Next we describe two experiments involving learning in the virtual world of *Desert Survival*. In particular, we are interested in the impact of attention window size A on the effectiveness of a mechanism similar to the psychological concept of Episodic

Memory.

In the second *Desert Survival* experiment, we endow the test Sheik with the capability of adapting α and β by evaluating past decisions and associating them with current inputs. He learns which values perform optimally for each camel depending on the circumstances (e.g., number of oases in the window, and how far the camel is located from the high saliency oases), and the parameters α and β are stored in memory as functions of the distances d_{ik} and the saliences s_k. Since learning proceeds by storing and evaluating snapshots of past *events*, it corresponds to Episodic Memory [3].

More specifically, the adaptive Sheik learns by associating the set of input parameters $\mathbf{u} = (s_1, \ldots, s_K; d_{i1}, \ldots, d_{iK})$, where K is the total number of oases within the window (K is roughly proportional to A^2), with a pair of output parameters (α, β) chosen from the sets $\alpha \in \{0.005, 0.1\}$, $\beta \in \{2, 10\}$. The effectiveness of the output parameters (whether they increase or decrease the well-being of the Sheik) is expressed by two "evaluation" parameters r_α, r_β, which are computed from the Sheik performance within a few time steps following the action taken. The parameter r_α encourages more aggression (larger α) if no life losses occured as a result of the action currently under evaluation. The parameter r_β encourages more demanding camel assignments (smaller β) if no territorial gains were obtained, and better life preservation if the territory improved during the monitoring period.

The learning procedure is as follows: let $\mathbf{u}_{\text{new}}$ denote the vector corresponding to the camel that the Sheik is currently considering, and let $\mathbf{u}_{\text{NN}}$ denote the nearest neighbor situation stored in Memory, i.e. the entry with minimum distance $B = ||\mathbf{u}_{\text{new}} - \mathbf{u}_{\text{NN}}||^2$ from $\mathbf{u}_{\text{new}}$. If $B \leq B_1$, then α_{NN} and β_{NN} are used and $r_{\alpha,\text{NN}}$, $r_{\beta,\text{NN}}$ are re-evaluated; if $B_1 < B \leq B_2$, α_{NN} and β_{NN} are still used but a new memory entry with coordinates $\mathbf{u}_{\text{new}}$ is created and evaluated; if $B > B_2$, the values of α, β are selected at random (*i.e.* the nearest neighbor is not used), and a new entry is again created and evaluated. We call this technique "reinforced Nearest Neighbor" (rNN) learning. A simplified theoretical model of rNN learning is given in sections 7.3.2-7.3.6. The parameters used in the experiments are $B_1 = 10$ and $B_2 = 25$. Memory

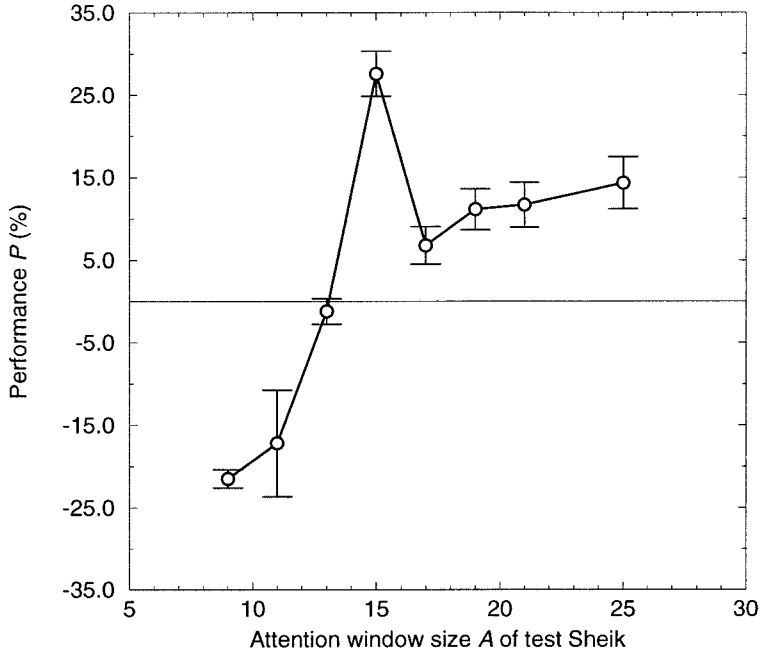


Figure 7.3: Performance P of the test Sheik who can learn episodically, and whose attentional window size A is varied against a non-learning control Sheik who uses the algorithm of (7.2) with a constant attentional window size $A = 15$.

access costs 1 *memory* token per search to the test Sheik, and he is given 1,875 memory tokens at every time step. There is no computational overhead assigned to learning.

The test Sheik, with variable window size, is paired against a non-learning control Sheik who applies the navigational algorithm (eq. 7.2) with fixed $\alpha = 0.1$ and $\beta = 1.0$ and fixed window size $A = 15$. These experiments last 200 virtual days per trial. The longer duration compared to the experiments of section 7.1.2 is necessary in order to allow the learning algorithm to reach saturation. See also section 7.3.6.

Two basic effects govern how P (defined as in section 7.1.2, eq. 7.3) changes with increasing window size (Figure 7.3). As A of the test Sheik increases, his memory stores a larger and larger fraction of the entire system, and his performance improves. Counteracting this trend, however, is the fact that evaluating all the associated probabilities for the camels and oases within this window takes up precious time and, as A increases, an increasingly smaller fraction of camels within the window become updated, decreasing performance. Thus, P peaks sharply at $A = 15$.

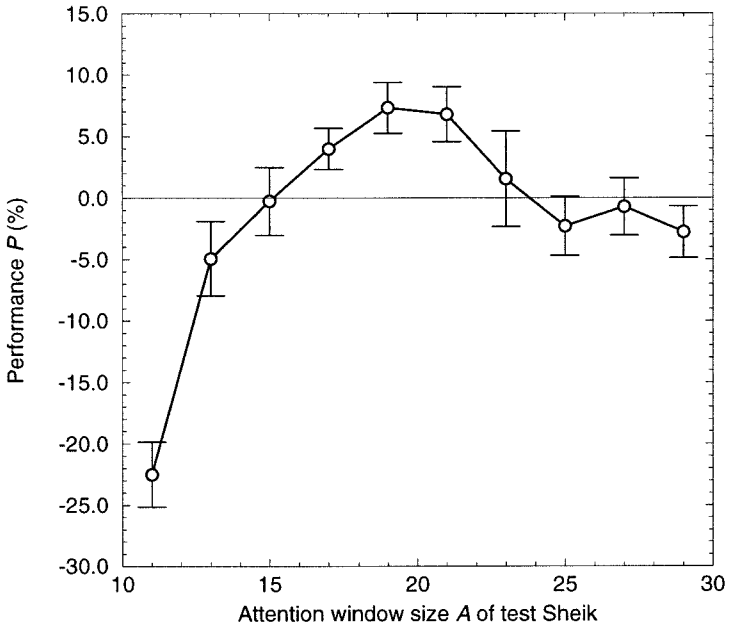


Figure 7.4: Performance P of the test Sheik, with variable attentional window size A, against a control Sheik, with fixed attentional window $A = 15$. Both Sheiks have episodic learning enabled.

In the final third experiment, both test and control Sheiks are allowed to adapt as described before; the only difference between them is that the size of the attentional window of the control Sheik is fixed at $A = 15$. In this experiment, memory search is free, since it is available to both Sheiks. Trials also last 200 virtual days each. As Figure 7.4 indicates, performance improves if the test Sheik has a larger window than the control Sheik, but only up to a point ($15 < A \leq 23$). Beyond this value, the time needed to update all camels with new sets of p_{ik}'s, that is with new marching orders, exceeds the total time allocated to each player. In this regime, the control Sheik has less information available for planning, but he is able to put it to more efficient use.

Concluding the discussion of the *Desert Survival* game, it is worth noting that an explicit, globally optimal strategy is not known. It is possible that a Nash equilibrium [158, 159, 160] exists, where the two opponents respond optimally to each other's strategy. If we eliminated the distinction between test and control, we stipulate that, aided by their attentional mechanisms, they would evolve (i.e., adaptively learn) their α and β in accordance with game-theoretic adaptive strategy models[4] [161], trying

[4]The parameters α, β determine the payoff matrices in our game.

to dynamically approximate the Nash equilibrium. In an experimental situation like the third simulation that we just described above, with the test Sheik having a non-optimal A, he should fail to adapt his strategy according to the demands of the position, because of information processing limitations (i.e., either too little or too much information, as explained above); hence, he would perform worse than the control Sheik. Experimental verification of this claim is made difficult by the fact that the Nash equilibrium (or even its existence) in *Desert Survival* is not known; yet, the claim presents an interesting topic of research, perhaps using a different complex game, designed so that a Nash equilibrium can be firmly established.

7.2 A biologically inspired computation model

Organisms living in time-varying, multi-dimensional environments face the problem of responding to real-time situations, and constructing viable plans for the near- or long-term future. At the same time, they must maintain maximum efficiency in computational apparatus and power consumption. We stipulate that advanced organisms (e.g., humans and primates in general) manage so by using the mechanism of awareness.

An enormous amount of variables describe the state of a complex organism at any given moment. On the one hand, sensors produce visual, auditory, tactile, and olfactory signals. For example, if we count the number of rods and cones in the retinas, we find that the dimension of the space on which the visual system alone can be described is, at a first glance, 250×10^6. On the other hand, we also know that sensory experiences are continuously stored as memories, which are recalled to aid the formulation of responses. Since each sensory event is described in a high dimensional space, the complexity of storing many of them, and in addition their sequences and associations, is clearly intractable.

Yet, in every day life, we do not sense the presence of so many variables in our behavior; indeed, it is very unlikely that we use all of them. If we did, the 10^{12} neurons in our nervous system would not suffice to exhaust the combinatorial explosion

of possibilities even for simple tasks, such as picking up a pencil, not to mention abstract reasoning and long-term planning. The hierarchical structure of the nervous system (probably a by-product of evolution [162, 163]) helps to organize cognitive tasks in systematic fashion [164]. For example, plans and intentions formed at the pre-frontal cortex of the brain elicit responses in the motor cortex, cerebellum, and basal ganglia, which in turn instruct the limbs to perform simple functions (such as "grabbing"). The peripheral knowledge about automated tasks is called Procedural Memory. Similarly, sensory inputs are not simultaneously processed at their entirety, but rather in a layered fashion, which has been studied most extensively in the visual system [165, 166]. Visual processing starts with the detection of very simple features, such as horizontal and vertical line segments, in the first layer (striate cortex, or area V1). In higher layers, more complicated tasks are performed, such as detection of object motion in area V4, and so on [165].

The results of the lower levels of sensory processing are sorted for significance with the aid of the mechanism of Attention [155, 156, 157]. The key function of Attention is to detect the most salient features in the input space [154, 167], thus restricting the amount of information to be processed by higher layers. The decision about saliency is either automatic ("bottom-up"), which serves to detect novelty and potential profit or danger in the environment; or deliberate ("top-down"), which serves to concentrate resources on a particular goal [168]. This filtering process is sometimes described as a "window" (or "spotlight") in sensory space. The result of attentional processing, or, in other words, what a human is "aware of" at any given moment, most often determines the human's subsequent response. Plausibly, therefore, Attention is Nature's solution to the dilemma of efficient biological computation with limited resources. Crick and Koch [148, 149] proposed that the function of awareness is

> "to produce the best current interpretation of the visual scene, in the light of past experience either of ourselves or of our ancestors (embodied in our genes), and to make it available, for a sufficient time, to the parts of the brain that contemplate, plan and execute voluntary motor outputs (of one sort or another)."

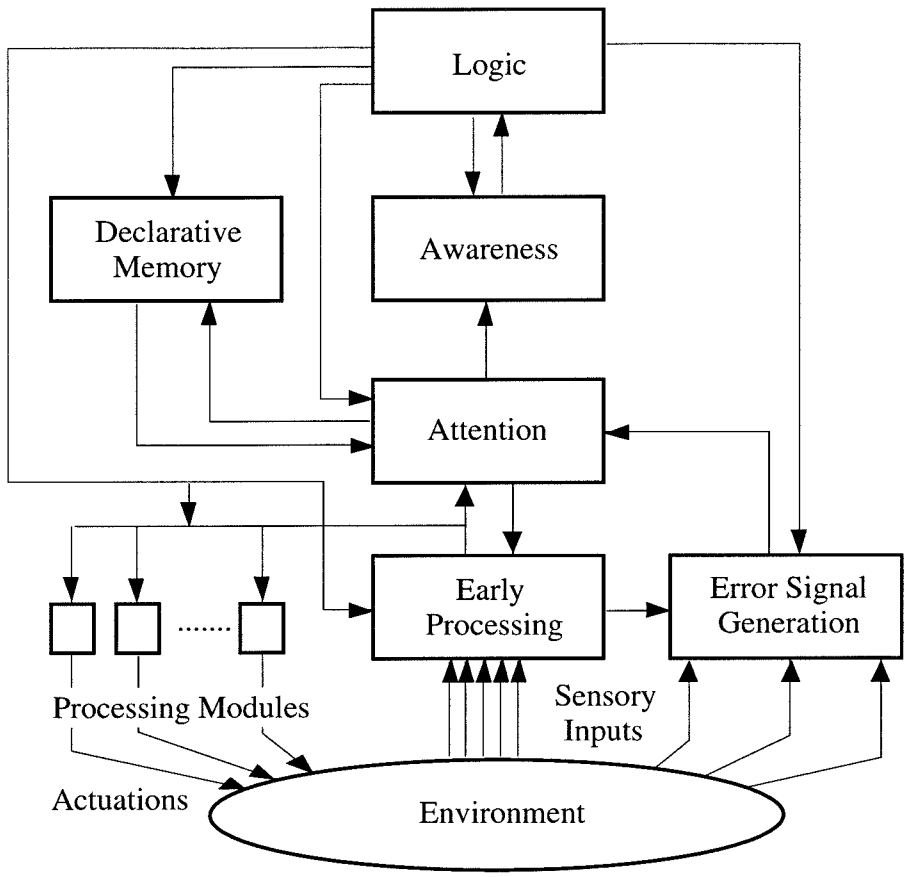


Figure 7.5: Block diagram of a computational architecture which is capable of forming efficient dynamic representations of the environment in real time.

The principle of limiting computational resources to the most significant subset of the input space at any given time was applied successfully to the *Desert Survival* game, as we saw in the previous section. A more general-purpose computational architecture, motivated by the same principles, is shown in Figure 7.5. It consists of two computational paradigms combined and running in parallel. The low-level modules are the Sensors, Early Processing, Error Module, and Actuators. Under normal conditions, this portion of the system operates autonomously, quickly responding to sensory inputs to perform familiar tasks. The higher-level system, comprised of Logic (a powerful CPU) and Declarative Memory [3, 9, 10, 169, 170, 171], implements emergency handling and long-term planning, by issuing command signals to the Actuators directly and configuration signals to all other modules. This portion of the system is

reminiscent of a von-Neumann architecture, except its inputs and outputs are routed via the low-level computing structure.

The Attentional module acts as a central information exchange. It selects the most salient subset of the pre-processed sensory inputs and then submits search requests to the Declarative Memory, in order to retrieve the most relevant past recollections. In turn, the memory returns a "memory window" of relevant facts. The memory should be organized in hierarchically to facilitate the search (see section 7.3.1). The results of sensory and memory window searches are fused to produce a compact representation, labeled Awareness. This is made accessible to the Logic unit, which then makes executive decisions in real-time. In this view, Awareness can be thought of as the state of the cache memory of the Logic, containing the information that is currently of most relevance to the system. In a world of unlimited computational resources, the system's Logic should have access to all information collected by the cooperating agents (the camels) for optimal behavior. In *Desert Survival*, Awareness represents the compromise needed to achieve better than chance performance in real-time (albeit most likely not optimal in a global sense) by providing the central processor with the most relevant information only.

The correspondence between the modules of Figure 7.5 and brain anatomy is well defined in some cases. Referring to the visual system, Early Processing is performed at the early levels of visual cortex (V1 through V4, with some IT functions possibly added in). The Actuators correspond to the motor cortex, basal ganglia and cerebellum, which control and coordinate movements and synchronize input/output from the limbs. Procedural memories are maintained locally in these modules [4]. Attention is believed to reside in the posterior parietal cortex (PPC) [172], although neuronal recordings indicate that visual responses are modulated by activity from most visual cortical areas [173, 174, 175, 176]. PPC perhaps acts as interface between attention and intention [17], equivalent to the connection between Logic and Attention in Figure 7.5. Logic and long-term planning in the brain are usually identified with the pre-frontal cortical areas [177].

By contrast, episodic and semantic (i.e., declarative) memory storage are not well

localized in the brain [3]; however, the hippocampus appears to be a switch controlling the passage of explicit (conscious) experiences into permanent storage. For instance, hippocampus is activated when explicit relational memories are formed [15, 178, 179, 180], but not if storage is implicit [6]. The neural correlate of awareness is not known, although several hypotheses have been made [181, 182, 183, 184, 185]. In the computational model, this problem is bypassed by defining the Awareness module as a cache memory, the contents of which are dynamically defined by the competition of inputs from Attention, Memory, and directives from Logic.

Returning to the model of *Desert Survival*, the camels perform simultaneously the functions of Sensors, Actuators, and Early Processing. The results of sensory processing produce the map of the desert, as in Figure 7.1. The actuations are water trading and oasis take-over. The Error Module calculates the set of saliences s_k and averages them over a running $A \times A$ window. The Attentional mechanism implements the selection of the most salient region in the desert and brings a compacted representation of the information about camels located inside the window (i.e., the saliences s_k and the distances d_{ik}) into the Awareness module. The Sheik corresponds to the Logic, which processes the Awareness information according to its own rules (the pre-programmed values of α, β, and eq. 7.2) to produce instructions (the p_{ik}'s) for the agents. If episodic learning is enabled, it re-configures the Logic at every step by providing values for α and β dependent on the Awareness information (the s_k's and d_{ik}'s). In turn, the Logic reconfigures memory by calculating the evaluation parameters r_α, r_β (see section 7.1.3).

7.3 Memory organization in the awareness model

7.3.1 Memory hierarchy

Taking a closer look at the awareness-based computational architecture, Figure 7.5, we observe that, in accordance to the biological evidence, there are two types of memory: procedural and declarative. The Procedural Memory is incorporated in the

Processing Modules themselves. This memory provides the necessary information for performing the specialized tasks that each module performs. Procedural Memory is accumulated through learning using supervised training algorithms with the help of the error signals. The pathway from the sensory input through the Early Processing leading to the Processing Modules can be thought of as a multi-layer neural network that can be trained with techniques such as back-propagation and variants. The Procedural Memory is then the set of adaptable weights in these networks.

The Declarative Memory is a large mass-memory that stores information which can be centrally accessed. This memory interacts with the rest of the system through command signals it receives from the Logic module and a two-way interaction pathway with the Attention module. The command signal from Logic can be thought of as an address signal that defines an accessible window or windows in the memory onto which the pathway between Declarative Memory and Attention operates. For instance, if the Awareness module indicates to the Logic that a face is present in the input visual scene, then the Logic elicits the part of the memory where faces are stored to come to the forefront. The pathway to the Attention module is used for associative recall. A pattern received from the Early Processing module is routed by Attention to the Declarative Memory, where it is compared to the contents of the accessible window. Once a match is found, the result is reported back to the Attention module. The data that is returned can be a compact representation, such as the address of the item that yielded a match, or a more detailed description of the same item. The Attention passes on this result to the Awareness and therefore makes the information (combined with the filtered input – see section 7.2) available to the Logic. The Logic then plans its next action which may include repositioning of the accessible window. The new information to the Attention box from the sensory inputs and leftover information received from memory are then compared with the new accessible window. In this way complex tasks requiring multiple memory access steps can be executed.

The organization of the Declarative Memory is both hierarchical and clustered. The hierarchical organization allows coarse categorization of inputs. Shifting the accessible window allows the system to zoom-in to make finer assignments. For in-

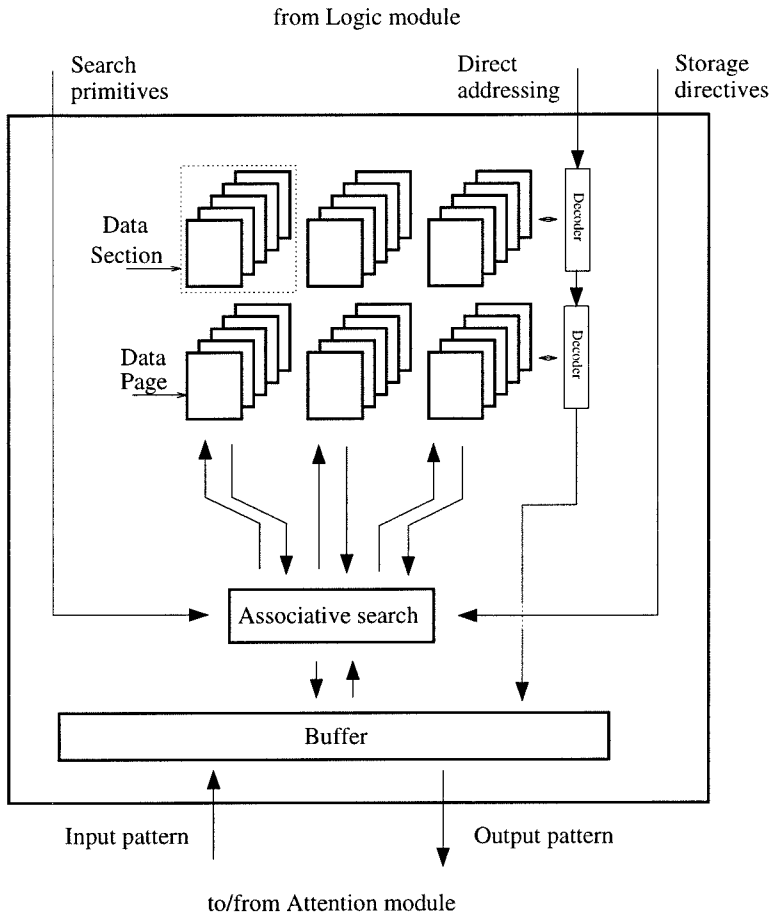


Figure 7.6: Hierarchical organization of the Declarative Memory.

stance, the top level of the memory can contain information that classifies an object as a face; lower levels store specific faces of familiar individuals. The clustered organization forces similar objects to be stored contiguously. This makes the search through an accessible window an efficient operation.

The database contained in the Declarative Memory is created through a learning process. If an object does not yield a sufficiently strong match, it is then added to the accessible window closest to the location that yielded the best match (the next few sections provide a detailed description). This way clustering is achieved. Data that need to be transferred from the Declarative Memory to the Processing Modules are routed in a compact representation through the Attention-Awareness-Logic pathway.

A generic design for the Declarative Memory is given in Figure 7.6. In the con-

text of holographic memories, we can imagine "data pages" as individual holograms, e.g., angle- or shift-multiplexed, whereas a section is a fractal row (recorded using peristrophic or fractal-angle or fractal-shift multiplexing, for example). Multiple sections are stored in spatially separated locations (space multiplexing). Many different schemes are possible, for example the compact module of section 6.3.3. Another possibility is recording the pages in the time-domain (see section 6.3.4), and using fractals for the sections. Notwithstanding the practical problem of capturing and processing the reconstructions (typically time-multiplexed holograms come out as ultra-short pulses at a rate of GHz), this organization may be appropriate for storing time sequences. Readout can be performed in one of two modes: Data requested by the Logic can be retrieved both directly (by-address) or associatively (by-content), whereas Attention communicates with Memory exclusively associatively.

The "memory window" is defined for associative search by the competition of Logic and Memory, by looking at a fraction of the correlations. Awareness can be thought of as occurring at multiple levels if there is a fixed number of detectors looking at the correlators. If the window "zooms out," a bigger portion of the Memory becomes visible but at a lower resolution. The degree of zoom and the location of the window are controlled by input saliency through Attention (similar to "bottom-up" attention) and by Logic when high level operations can be performed safely while ignoring the environment ("top-down" attention).

7.3.2 Learning uncertain environments

In section 7.1.3, we described episodic learning (or learning-by-experience) in the *Desert Survival* game, and showed how learning efficiency versus a non-learning (Figure 7.3) or learning (Figure 7.4) opponent is affected by the attentional window size A. For the purposes of the present discussion, the virtual desert is a *random environment*, meaning an environment where a specific action in response to a specific situation (captured by the input signals) produces a random result, sometimes beneficial and sometimes harmful. In a random environment, the purpose of learning is

to determine the probability distributions of the results conditioned on actions and situations, and the optimal (in Bayesian sense) response strategy. Our model is a simplification of the actual scheme used in *Desert Survival*, yet it captures the associated trade-offs well.

Lacking a model for the random environment, the learning algorithm must necessarily start with a randomized strategy; i.e., the algorithm is initialized with each possible decision being equally probable in response to any signal. As time evolves, the algorithm should evaluate the results of past decisions to find the one statistically most profitable as fast as possible, and segment the signal space by finding the boundaries where the optimal decision changes. We are interested both in the degree of approximation to the Bayesian strategy that the learning scheme can achieve, as well as the amount of time required to do so.

More specifically, let $\mathcal{S} \subseteq \mathrm{R}^d$ denote the signal space[5], and

$$p(\mathbf{x}), \quad \mathbf{x} \in \mathcal{S}$$

the a-priori probability that $\mathbf{x}$ occurs. Let

$$Y = \{y_1, y_2, \ldots, y_D\}$$

denote the set of decisions. We assume that functions $r(\mathbf{x}; y) \in \mathrm{R}$ and $q(\mathbf{x}; y; r(\mathbf{x}; y)) \in [0, 1]$ exist such that, $\forall \mathbf{x} \in \mathcal{S}$ and $y \in Y$, return r occurs with probability q if action y is taken in response to signal $\mathbf{x}$. All the distributions are assumed to have at worst a countable number of step discontinuities. Let r be a utility function with $r \geq 0$ denoting gain and $r < 0$ denoting loss. The objective of the algorithm can now be restated as finding action $y_{\mathrm{opt}}(\mathbf{x})$ which optimizes the expected return

$$\tilde{r} = \mathrm{EV}\{r \mid \mathbf{x}, y_{\mathrm{opt}}(\mathbf{x})\}, \tag{7.4}$$

[5]In a simple system, $\mathcal{S}$ is simply the space spanned by the input signals. In the computational model of Figure 7.5, $\mathcal{S}$ is the space where the contents of the Awareness module take their values from.

where EV denotes expectation value over the distributions p, q.

7.3.3 The rNN algorithm

The reinforced Nearest Neighbor (rNN) algorithm is a variant of the nearest-neighbor algorithm that uses reinforcement, since no labelling of the data is available. In particular, the algorithm constructs a memory space $\mathcal{M}$ homologous to the sensory space $\mathcal{S}$. Initially, $\mathcal{M}$ is empty. When a signal $\mathbf{x} \in \mathcal{S}$ arrives (drawn from the distribution p), the algorithm checks a d-dimensional sphere of radius b around $\mathbf{x} \in \mathcal{M}$. We will refer to b as the proximity radius. If no memory is found within that sphere, a random decision is drawn from Y according to some distribution s. If a memory $\mathbf{x}' \in \mathcal{M}$ is found within distance b from $\mathbf{x}$, then the algorithm checks the decision y stored in $\mathbf{x}'$ and its history of returns r. If the history shows that y is profitable according to the utility criterion, then y is repeated in response to $\mathbf{x}$, otherwise another action is drawn from Y according to a rejection rule. In the case when Y contains only $D = 2$ actions, the rejection rule is to take simply the alternative action. In addition, if the distance of "recollection" $\mathbf{x}'$ from $\mathbf{x}$ is more than $b/2$, then a new memory element is created at position $\mathbf{x} \in \mathcal{M}$. Clearly, this is a simplification of the *Desert Survival* model (section 7.1.3) with $B_1 = B_2 = b$.

We seek to understand reinforcement rules for the rNN algorithm, and determine the optimal learning rate, expressed by the proximity radius b. Indeed, the efficiency and accuracy of the process of filling up the memory space are conflicting as we will show in detail later. If b is large, then the memory space is filled up fast; however, the decision boundary is blurred. In the opposite case of small b, the decision boundary may be determined very accurately; however, random decisions are taken very often, since the space is filled very slowly. Therefore, each one of these extremes may result in ruin with high probability.

To keep the analysis simple, we will restrict ourselves to the simple $2 \times 2 \times 2$ configuration shown in Figure 7.7. The two-dimensional parameter space is segmented in two regions by a straight decision boundary. In addition, there are only two possible

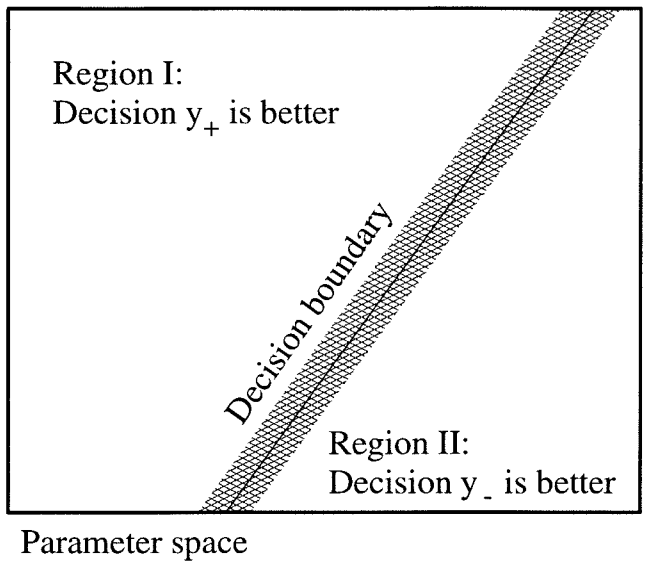


Figure 7.7: Description of the decision problem in the $2 \times 2 \times 2$ configuration (two-dimensional space, two possible decisions, two possible results).

actions: $Y = \{y_+, y_-\}$, and two possible returns: $r \in \{+1, -1\}$. The causal relation between actions and results is the following:

$$y_+ \text{ causes } \begin{cases} r = +1 \text{ with probability } \dfrac{1+\varepsilon}{2} \\ r = -1 \text{ with probability } \dfrac{1-\varepsilon}{2} \end{cases}, \tag{7.5}$$

$$y_- \text{ causes } \begin{cases} r = +1 \text{ with probability } \dfrac{1-\varepsilon}{2} \\ r = -1 \text{ with probability } \dfrac{1+\varepsilon}{2} \end{cases}, \tag{7.6}$$

where $\varepsilon \in [0, 1]$ is the certainty parameter ($\varepsilon = 1$ makes the problem deterministic). Clearly action y_+ is "preferrable," since it produces positive r "more often." In fact it is easy to see that the optimal (Bayesian) strategy consists of picking action y_+ *always*, which we will denote as

$$\psi_+^{\text{opt}} = \Pr(y = y_+) = 1. \tag{7.7}$$

Then the probability of winning is

$$\phi_+^{\text{opt}} = \Pr(r = +1) = \frac{1+\varepsilon}{2}, \tag{7.8}$$

and the expected return is

$$\tilde{r} = \phi_+^{\text{opt}} \times (+1) + \left(1 - \phi_+^{\text{opt}}\right) \times (-1) = \varepsilon. \tag{7.9}$$

Generalizations such as piecewise linear or non-linear decision boundary, higher dimensional input space, and more than two possible decisions are straightforward, based on the framework presented in the subsequent sections and will not be presented here. Extention to the *Desert Survival* model (section 7.1.3) is also straightforward, and will be omitted.

7.3.4 Reinforcement algorithms

We seek to develop an algorithm for deciding what the optimal strategy is by observing the response of the environment. At time $t = 1$ the algorithm picks either one of $y_{+/-}$ with probability 1/2. Beyond that point, the algorithm should asymptotically (i.e., as $t \to +\infty$) achieve performance equal to ϕ_+^{opt}. We will present four algorithms, and show that only by looking back at the entire past is it possible to achieve Bayesian performance.

Algorithm A: Single-step look-back

At each step t the system examines the action and result $r(t-1)$ incurred at the latest step. If $r(t-1) = +1$, then the system takes the same action $[y(t) = y(t-1)]$, otherwise the latest action is reversed. We will now prove the following

Proposition: The expected performance of algorithm A at step t ($t \geq 2$) is

$$\phi_+(t) = \frac{1+\varepsilon^2}{2}. \tag{7.10}$$

Proof: From the definition of algorithm A we have:

$$\begin{aligned}
\psi_+(t) &= \Pr\Big([y(t-1) = y_+ \text{ and } r(t-1) = +1] \text{ or} \\
&\quad [y(t-1) = y_+ \text{ and } r(t-1) = +1]\Big) \\
&= \Pr\Big(r(t-1) = +1 \mid y(t-1) = y_+\Big) \Pr\Big(y(t-1) = y_+\Big) + \\
&\quad \Pr\Big(r(t-1) = -1 \mid y(t-1) = y_-\Big) \Pr\Big(y(t-1) = y_-\Big) \\
&= \frac{1+\varepsilon}{2}\psi_+(t-1) + \frac{1+\varepsilon}{2}\psi_-(t-1) \\
&= [\psi_+(t-1) + \psi_-(t-1)]\frac{1+\varepsilon}{2} = \frac{1+\varepsilon}{2} \\
\phi_+(t) &= \Pr\Big(r(t) = +1 \mid y(t) = y_+\Big) \Pr\Big(y(t) = y_+\Big) + \\
&\quad \Pr\Big(r(t) = +1 \mid y(t) = y_-\Big) \Pr\Big(y(t) = y_-\Big) \\
&= \psi_+(t)\frac{1+\varepsilon}{2} + \psi_-(t-1)\frac{1-\varepsilon}{2} \\
&= \left(\frac{1+\varepsilon}{2}\right)^2 + \left(\frac{1-\varepsilon}{2}\right)^2 = \frac{1+\varepsilon^2}{2}. \qquad \triangle
\end{aligned}$$

Therefore, algorithm A fails to satisfy the requirement of asymptotic convergence to ϕ_+^{opt}. The expected gain is $\tilde{r} = \varepsilon^2$.

Algorithm B: Linear randomization

At time t, the system examines the results at all previous steps. Let's denote by m the number of "favorable" outcomes, i.e., occurrences of y_+ resulting in $r = +1$ and y_- resulting in $r = -1$. The action $y(t)$ at step t ($t \geq 2$) is chosen according to the following rule:

$$y(t) = \begin{cases} y_+ & \text{with probability } \dfrac{m}{t-1} \\ y_- & \text{with probability } 1 - \dfrac{m}{t-1} \end{cases} \tag{7.11}$$

This randomized strategy seems more reasonable than algorithm A; however, it turns out to yield the same sub-optimal result:

Proposition: The expected performance of algorithm B at step t ($t \geq 2$) is

$$\phi_+(t) = \frac{1+\varepsilon^2}{2}. \tag{7.12}$$

Proof: The probability of m "favorable" outcomes occurring in t steps is given by the Bernoulli distribution

$$\psi_+^m(t+1) = \binom{t}{m} \left(\frac{1+\varepsilon}{2}\right)^m \left(\frac{1-\varepsilon}{2}\right)^{t-m}. \tag{7.13}$$

Therefore, the probability of taking the correct action y_+ at step $t+1$ is

$$\begin{aligned}
\psi_+(t+1) &= \sum_{m=1}^{t} \frac{m}{t} \binom{t}{m} \left(\frac{1+\varepsilon}{2}\right)^m \left(\frac{1-\varepsilon}{2}\right)^{t-m} \\
&= \frac{1+\varepsilon}{2} \sum_{m=1}^{n} \binom{t-1}{m-1} \left(\frac{1+\varepsilon}{2}\right)^{m-1} \left(\frac{1-\varepsilon}{2}\right)^{t-m} \\
&= \frac{1+\varepsilon}{2} \sum_{m=0}^{t-1} \binom{t-1}{m} \left(\frac{1+\varepsilon}{2}\right)^{m} \left(\frac{1-\varepsilon}{2}\right)^{t-1-m} \\
&= \frac{1+\varepsilon}{2} \left[\left(\frac{1+\varepsilon}{2}\right) + \left(\frac{1-\varepsilon}{2}\right)\right]^{t-1} = \frac{1+\varepsilon}{2}.
\end{aligned}$$

$\phi_+(t)$ is then calculated according to the last step of the proof for algorithm A. $\triangle$

Algorithm C: Infinite Majority vote

This is similar to algorithm B, except the decision follows a majority rule: y_+ is taken if $m > t/2$, and y_- is taken if $m < t/2$. If t is even and $m = t/2$, then either y_+ or y_- is picked randomly with equal probabilities $1/2$. The probability of having a majority of "favorable" occurrences at step $t+1$ ($t \geq 2$, even) is

$$\psi_+(t+1) = \sum_{m=\frac{t+1}{2}}^{t} \binom{t}{m} \left(\frac{1+\varepsilon}{2}\right)^m \left(\frac{1-\varepsilon}{2}\right)^{t-m}. \tag{7.14}$$

Random walk theory (section XI.3, p. 242 of [186]) states that as t increases, sooner or later the "favorable" occurrences m overtake the unfavorable ones, and therefore $\psi_+(t+1)$ tends to one as $t \to \infty$. This is verified in the plots of ψ_+, ϕ_+ given in Figure 7.8. From the plots we also infer that the convergence time is approximately inversely proportional to ε.

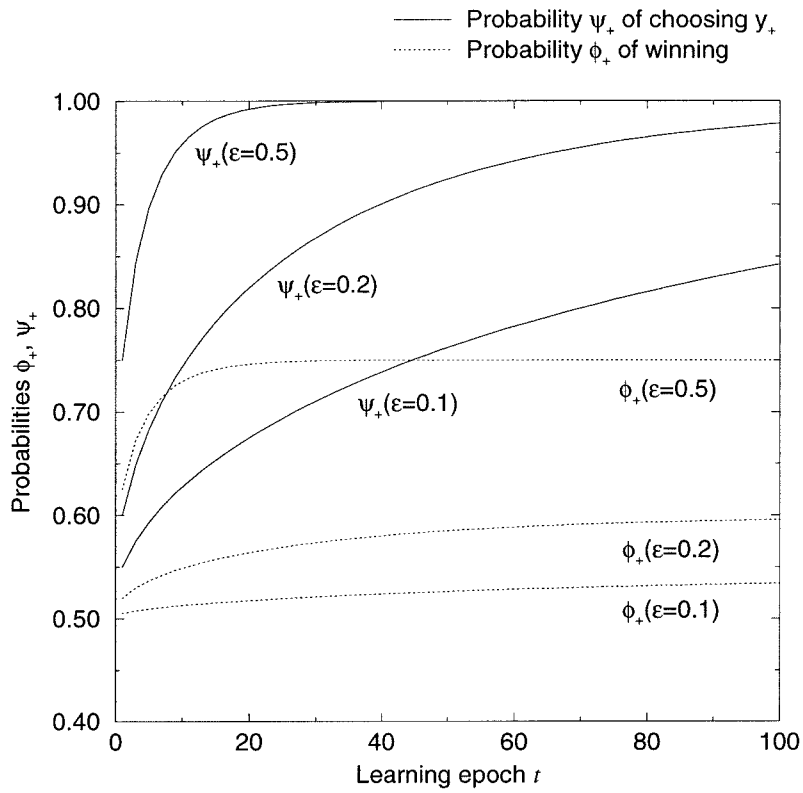


Figure 7.8: Convergence of the majority vote rule.

The evolution of ψ_+ can be approximated well by the following heuristic expression:

$$\psi_+(t) \approx 1 - \frac{1}{2}\exp\left\{-\kappa_1 t^{\zeta_1} - \kappa_2 t^{\zeta_2}\right\} \tag{7.15}$$

The parameters κ_1, ζ_1, κ_2, ζ_2 as functions of ε are determined by least-squares fit. The result is shown in Figure 7.9. With this approximation, the probability of winning at time t is

$$\phi_+(t) \approx \frac{1 + \varepsilon\left(1 - \exp\left\{-\kappa_1 t^{\zeta_1} - \kappa_2 t^{\zeta_2}\right\}\right)}{2}. \tag{7.16}$$

Algorithm D: Truncated Majority vote

Suppose that the memory does not have the capacity to perform a majority decision over the entire past, but only over the most recent M steps at any give time step. We call this algorithm Truncated Majority. During the first M time steps after the

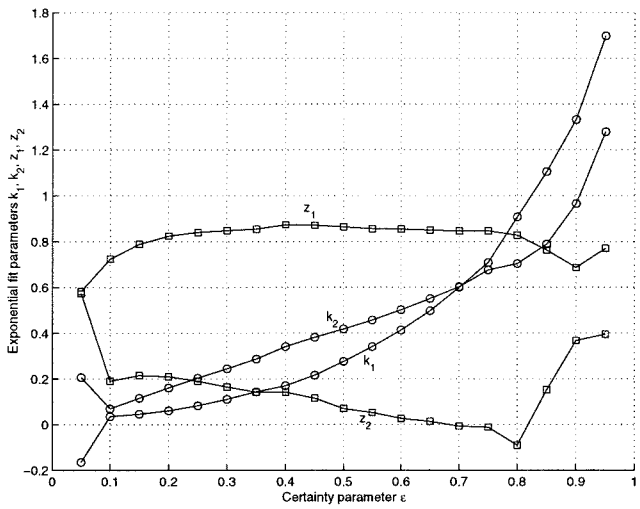


Figure 7.9: Parameters κ_1, ζ_1, κ_2, ζ_2 of the exponential fit (7.15) as functions of the certainty parameter ε.

original creation of the memory, the algorithm performs exactly as Algorithm C, i.e., (7.14) applies. However, for times $t > M$ the performance saturates and ψ_+ cannot increase beyond

$$\psi_+(t \geq M) = \psi_+(M) = \sum_{m=\frac{M+1}{2}}^{M} \binom{M}{m} \left(\frac{1+\varepsilon}{2}\right)^m \left(\frac{1-\varepsilon}{2}\right)^{t-m}. \tag{7.17}$$

For instance, in the examples of Figure 7.8, with $M = 60$ ψ_+ would be limited to ≈ 0.77 for $\varepsilon = 0.1$ and ≈ 0.92 for $\varepsilon = 0.2$, never approaching 1 as in the unforgetful case. However, if the environment changes with a time scale approximately equal to M, then forgetting may be beneficial because it helps the system adapt faster. This was certainly the case in *Desert Survival* where forgetting was implicitly imposed by the constraints of real-time operation (see section 7.1.3).

7.3.5 Dynamics of memory occupancy

In this section we determine the time evolution of memory occupancy and the optimal learning rate, in other words the optimal proximity radius b_{opt}. As in the previous

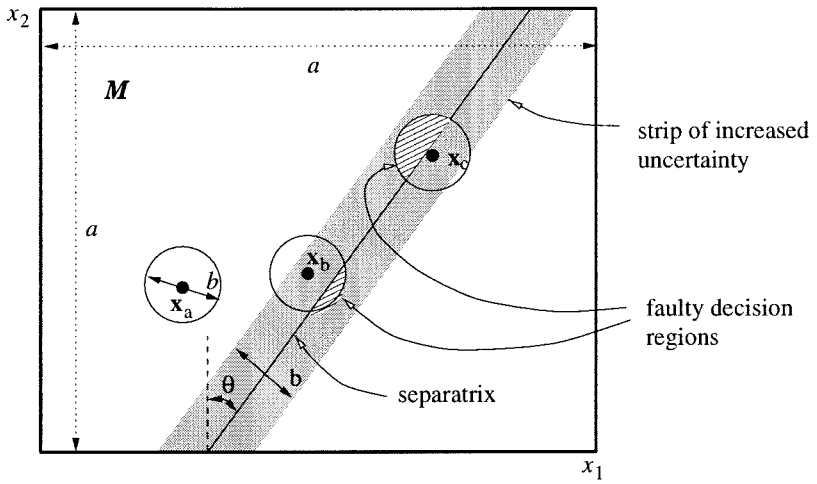


Figure 7.10: Increased uncertainty at the decision boundary.

section, we will restrict ourselves to the simple 2×2 case, and, moreover, we will use Algorithm C ("infinite majority vote") for reinforcement. The result depends on the a-priori distribution p which we will assume to be uniform in $[0, a]^2$. We break the calculation down into two steps: first we calculate the dependence of saturation performance (after each point has converged to $\psi_+ \approx 1$) on b; then we can calculate approximately the learning rate as function of b, and determine the ratio b/a that maximizes the learning rate as function of ε.

Saturation learning efficiency

Let $\mathcal{M} = [0, a] \times [0, a] \subset \mathrm{R}^2$ and let the separatrix be a straight line making angle θ with the x_2-axis (see Figure 7.10). Consider the three memorized points $\mathbf{x}_\mathrm{a}$, $\mathbf{x}_\mathrm{b}$, $\mathbf{x}_\mathrm{c}$ of Figure 7.10, and assume that reinforcement learning is performed according to Algorithm C (infinite majority rule) of the previous section. Also, assume that the action-result relationship to the left of the separatrix is given by (7.5-7.6), whereas to the right of the separatrix, the relationship is the same but with y_+ and y_- interchanged. Point $\mathbf{x}_\mathrm{a}$ will achieve optimal performance equal to $\tilde{r} = \varepsilon$. However, the memory located at $\mathbf{x}_\mathrm{b}$ will not perform that well, since the shaded part of the proximity circle falls on the opposite side of the decision boundary. The same is true for point $\mathbf{x}_\mathrm{c}$ and any other point falling inside the strip extending one proximity

distance around the separatrix. The evaluation of memories $\mathbf{x}_b$, $\mathbf{x}_c$ is made more difficult because the effective certainty decreases. If g, $|g| < b/2$, is the distance from the memory point $\mathbf{x}$ to the separatrix, then a simple geometrical calculation shows that the effective certainty is

$$\varepsilon_{\text{eff}}(g) = \varepsilon \left[1 - \frac{8}{\pi} \left(\frac{1}{2} \arccos \frac{2g}{b} - \frac{g}{b} \sqrt{1 - \left(\frac{2g}{b} \right)^2} \right) \right]. \tag{7.18}$$

By integrating over the possible distances $0 \leq g \leq b/2$, we find the effective certainty over the entire strip

$$\varepsilon_{\text{eff}}(\text{strip}) = \varepsilon \left(1 - \frac{4}{3\pi} \right). \tag{7.19}$$

Over the entire parameter space,

$$\varepsilon_{\text{eff}}(\mathcal{M}) = \varepsilon \left(1 - \frac{4b}{3\pi a \cos\theta} \right). \tag{7.20}$$

Therefore, the probability of winning decreases with b as

$$\phi_+(\infty) = \frac{1}{2} \left[1 + \varepsilon \left(1 - \frac{4b}{3\pi a \cos\theta} \right) \right]. \tag{7.21}$$

This result verifies the intuitive notion that the larger b is, the more unlikely it becomes that the reinforcement algorithm approximates the decision boundary correctly.

Learning rate

Proposition: The expected memory occupancy at time t ($t \geq 1$), i.e., the expected number of stored memories, is

$$n(t) = \frac{4a^2}{\pi b^2} \left[1 - \left(1 - \frac{\pi b^2}{4a^2} \right)^t \right] \tag{7.22}$$

Proof: At time $t = 1$, the occupancy is $n(1) = 1$. Let $n(t)$ be the occupancy at time t. Each disk occupies area $\pi b^2/4$; therefore, the probability that a new memory element

will occur inside the area occupied by the disks is

$$\frac{n(t)\pi b^2}{4a^2}.$$

It follows that the expected number of disks at the next time step is

$$n(t+1) = n(t)\frac{n(t)\pi b^2}{4a^2} + (n(t)+1)\left(1 - \frac{n(t)\pi b^2}{4a^2}\right) \tag{7.23}$$

$$= n(t)\left(1 - \frac{\pi b^2}{4a^2}\right) + 1. \tag{7.24}$$

By a simple induction we obtain that $n(t)$ is given for $t \geq 1$ by (7.22). $\triangle$

We will now show that the average learning progress at time t ($t \geq 1$), is, to good approximation,

$$\psi_+^{\text{eff}}(t) \approx \frac{1}{2}\left(1 - \frac{\pi b^2}{4a^2}\right)^t + \frac{\pi b^2}{4a^2}\sum_{k=0}^{t-1}\left(1 - \frac{\pi b^2}{4a^2}\right)^k \left(1 - \frac{1}{2}\exp\left\{-\kappa_1 t^{\zeta_1} - \kappa_2 t^{\zeta_2}\right\}\right). \tag{7.25}$$

Indeed, from the previous result we can derive that the expected number of memories added at step $t+1$ is

$$n(t+1) - n(t) = \left(1 - \frac{\pi b^2}{4a^2}\right)^t. \tag{7.26}$$

Therefore, when a new input appears, the probability that it will coincide with a memory of age k is

$$\frac{\pi b^2}{4a^2}\left(1 - \frac{\pi b^2}{4a^2}\right)^{k-1},$$

and the probability that it will not coincide with any existing memories is

$$1 - \sum_{k=1}^{t}\frac{\pi b^2}{4a^2}\left(1 - \frac{\pi b^2}{4a^2}\right)^{k-1} = \left(1 - \frac{\pi b^2}{4a^2}\right)^t.$$

In the former case, ψ_+ is given approximately by (7.15); in the latter, it is simply $1/2$ (randomized decision). Combining these results we obtain (7.25). Strictly speaking,

κ_1, ζ_1, κ_2, ζ_2 must be calculated separately for $\varepsilon_{\text{eff}}(g)$ to account for the edge effects at the decision boundary, and then (7.25) must be averaged over g to obtain the learning rate. However, to keep the analysis tractable, in the calculations we will use the κ_1, ζ_1, κ_2, ζ_2 corresponding to $\varepsilon_{\text{eff}}(\mathcal{M})$ (eq. 7.20). The probability of winning as time progresses is, therefore,

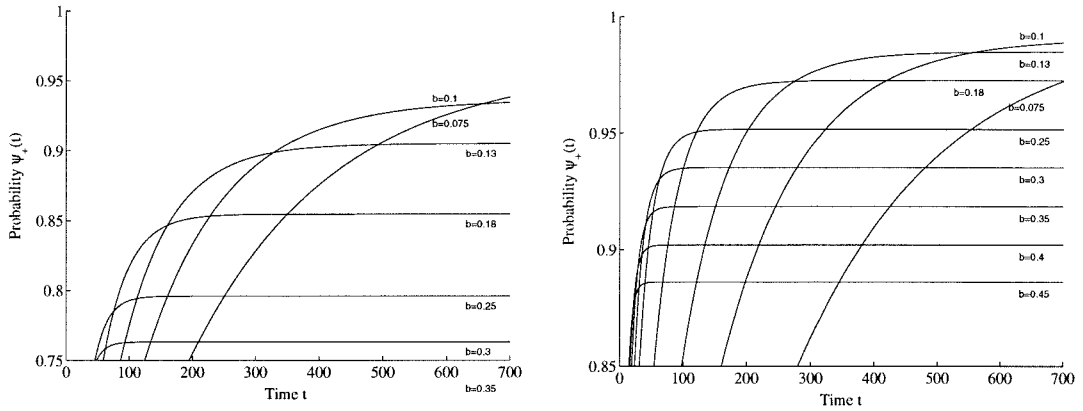
$$\phi_+^{\text{eff}}(t) = \frac{1}{2} + \varepsilon_{\text{eff}}(\mathcal{M}) \left(\psi_+^{\text{eff}}(t) - \frac{1}{2} \right). \tag{7.27}$$

Generalizations

Several generalizations are possible and straightforward in terms of the framework given above. Generalization to higher dimensions is straightforward using existing analyses of the problem of intersecting spheres inside a cube in d dimensions (e.g., [187]). The case of a piecewise linear or non-linear separatrix is treated easily with the same theory presented above, by modifying the integration at the boundary. Other interesting possibilities arise if the a-priori distribution is other than uniform (e.g., Gaussian), or non-stationary (in the latter case forgetting, as in Algorithm D, would have to be employed).

7.3.6 Optimization of the learning rate

Figure 7.11 shows the dynamics (7.27) of rNN learning for two different values of the certainty parameter ε. It is clear from the figures that the optimum b depends on the time t when we want to achieve the goal of optimum performance. For example, for the case $\varepsilon = 0.2$ (Figure 7.11a), if we can afford to wait until $t = 700$ to achieve maximum performance, then $b = 0.075$ is the best choice. However, the price to pay for this choise is sub-optimal performance for a long initial time period, $t < 650$. If the long wait has disastrous consequences, because of the losses that will incur during the long learning period, then one has to settle for a larger b (e.g., $b = 0.1$ or $b = 0.13$), which yields faster learning rate but also lower value of ψ_+ at saturation. Similar reasoning holds for $\varepsilon = 0.5$ (Figure 7.11b), except as ε increases, the performance of



(a) (b)

Figure 7.11: Memory dynamics: evolution of the probability ψ_+ of taking the correct decision with time t for different values of b, and (a) $\varepsilon = 0.2$, (b) $\varepsilon = 0.5$.

larger b improves because the residual errors at the boundary become less significant.

The conclusion from this discussion and the analysis of the previous sections for the simple $2 \times 2 \times 2$ case is that learning in an uncertain environment must be carefully tuned to the required time pace and learning efficiency. Systems with reasonable configurations manage to survive by performing well enough in time (although perhaps sub-optimally in the long run), whereas any other configuration would be eliminated through evolution.

7.4 Conclusions, discussion, and future extensions

Several cognitive functions, apart from those discussed herein, may be studied with the aid of the *Desert Survival* simulation. For example, inhibition/facilitation of return to the attentional focus may improve or deteriorate performance depending on various desert parameters. Prediction of the opponent is a significant feature of intelligence. Currently, each Sheik predicts the future by assuming that his counterpart is similar to himself. Possibly, Awareness may allow better performance by classifying the opponent into one of several categories based on his observed behavior, and then adjusting one's own strategy accordingly. Finally, forgetting deserves some more

consideration as a means of adaptation, as we mentioned in section 7.3.4.

Potential applications of the awareness-based approach to computation involve situations where machines are required to solve computationally overwhelming problems, yet problems which humans manage to solve satisfactorily with "sub-optimal" reasoning. Examples of applications are interactive automated office assistants, autonomous agents for hazardous industrial or extra-terrestrial environments, traffic control in metropolitan areas, air traffic control, and managing large physical plants (e.g. oil refineries, hospitals, aircraft carriers and so on).

As a specific application example with immediate commercial interest, consider an *Intelligent Building*. Referring to Figure 7.5, the Processing Modules correspond to face- and voice-recognition modules at the doors, cameras and PCs tracking the movement of humans inside the building, temperature and humidity sensors in every room, and so on. This vast data stream is transcribed into a detailed representation of the entire building. An attentional mechanism needs to select a subset of this data and generate a more compact representation—corresponding to the current contents of awareness—which is made available to the Logic unit (a central CPU). The error signals assign higher priority to events such as the detection of an intruder or a fire, compared to the routine identification of an employee walking through the front door. Many situations require access to Memory: for instance, comparing an unknown face against the faces of known criminals stored in a database or checking whether a high occupancy in one of the lecture rooms corresponds to a regularly scheduled seminar. A compact, dynamic and flexible "awareness" representation implemented on an inexpensive personal computer might be sufficient to help manage such a large-scale system. Alternatively, the contents of the Awareness module may be forwarded directly to a human operator, who in that case would be performing the function of the Logic module.

Another application example is computer chess. Like *Desert Survival*, chess requires no complicated sensory processing or actuation control algorithms. The inputs are the locations of at most 32 pieces (including those of the player and those of the opponent) on an 8×8 grid (the "chessboard"), and the response is a single move at

a time. Commercial chess computers typically calculate their response by evaluating an expanding tree of possibilities, starting from the current position and searching a depth of a few moves. Because of the nature of chess rules, the number of possibilities that result from this search grows exponentially with search depth. By contrast, the best human players usually evaluate explicitly very few possible future positions, however they are very selective in the move sequences they consider. It seems, therefore, that an attentional mechanism exists, which can be trained by experience[6], and suggests moves that should be considered as the most "reasonable" when the computational resources are limited (the thinking time in competitive chess is typically 2 hours for 40 moves; in a different format, "speed chess," only 5 minutes are allowed for the entire game). The applicability of the computational model of Figure 7.5 is now clear, although the details of the implementation are beyond the scope of this work. A chess computer built along the lines of the awareness-based model would be readily comparable to traditional architectures, in terms of a cost (or complexity) *vs.* performance quotient.

Many challenging computational tasks can be successfully attacked by concentrating sufficient computational resources. For example, IBM recently assembled a massively parallel computer to evaluate 200 million chess positions per second. This sufficed to overcome the problem of combinatorial explosion of possibilities with search depth, and the computer defeated the world champion[7]. If, however, resources are limited, as in most physical systems, a method is needed to compromise performance for efficiency in terms of power, size, time, and cost. The approach used in the *Desert Survival* game, inspired by the cognitive architecture of the primate brain, can provide one answer. It can lead to a novel computational paradigm appropriate for solving complex problems efficiently by taking decisions based on a compact representation of the most salient features of the environment and the relevant information from memory.

[6]Typically, untrained players try to consider *all* possibilities, to no avail.

[7]For information on Deep Blue and the match against Grandmaster Garry K. Kasparov, consult `http://www.chess.ibm.com`, and `http://www.rs6000.ibm.com`.

Bibliography

[1] T. Ribot, *Les maladies de la memoire*, Germer Ballaire, Paris, 1881.

[2] H. Ebbinghaus, *Über das Gedachtnis*, Duncker & Humblot, Leipzig, 1885.

[3] L. R. Squire, *Memory and brain*, Oxford University Press, 1987.

[4] M. Mishkin and T. Appenzeller, "The anatomy of memory," *Sci. Am.*, 256:80–89, 1987.

[5] A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio, "Deciding advantageously before knowing the advantageous strategy," *Science*, 275(5304):1293–1295, 1997.

[6] G. S. Berns, J. D. Cohen, and M. A. Mintun, "Brain regions responsive to novelty in the absence of awareness," *Science*, 276:1272–1275, 1997.

[7] S. S. Korsakoff, "Disturbance of psychic function in alcoholic paralysis and its relation to the disturbance of the psychic sphere in multiple neuritis of nonalcoholic origin," *Vestn. Psychiatrii*, 4(2), 1887.

[8] W. James, *Principles of phsychology*, Holt, New York, 1890.

[9] L. R. Squire, "Declarative and non-declarative memory: multiple brain systems supporting learning and memory," in D. L. Schacter and E. Tulving, editors, *Memory Systems 1994*, pages 203–232, MIT Press, 1994.

[10] E. Tulving, "Concepts of human memory," in L. R. Squire, N. M. Weinberger, G. Lynch, and J. L. McGaugh, editors, *Memory: Organization and locus of change*, pages 3–32, Oxford University Press, 1991.

[11] D. L. Schacter and E. Tulving, "What are the memory systems of 1994?," in D. L. Schacter and E. Tulving, editors, *Memory Systems 1994*, pages 1–38, MIT Press, 1994.

[12] M. Coltheart, "Iconic memory," *Phil. Trans. Royal Soc., London*, B302:183–294, 1983.

[13] G. Sperling, "The information available in brief visual presentations," *Psychol. Monogr.*, 60, 1960.

[14] H. Eichenbaum, N. J. Cohen, T. Otto, and C. Wible, "Memory representation in the hippocampus: functional domain and functional organization," in L. R. Squire, N. M. Weinberger, G. Lynch, and J. L. McGaugh, editors, *Memory: Organization and locus of change*, pages 163–204, Oxford University Press, 1991.

[15] H. Eichenbaum, "The hippocampal system and declarative memory in humans and animals: experimental analysis and historical origins," in D. L. Schacter and E. Tulving, editors, *Memory Systems 1994*, pages 147–202, MIT Press, 1994.

[16] E. Tulving, S. Kapur, H. J. Markowitsch, F. I. M. Craik, and R. Habib, "Neuroanatomical correlates of retrieval in episodic memory: auditory sentence recognition," *Proc. Natl. Acad. Sci.* (USA), 91(6):2012–2015, 1994.

[17] L. H. Snyder, A. P. Batista, and R. A. Andersen, "Coding of intention in the posterior parietal cortex," *Nature*, 386(6621):167–170, 1997.

[18] A. Karni and D. Sagi, "Where practice makes perfect in feature-discrimination – evidence for primary visual cortex plasticity," *Proc. Natl. Acad. Sci. USA*, 88(11):4966–4970, 1991.

[19] A. Karni and D. Sagi, "The time-course of learning a visual skill," *Nature*, 365(6443):250–252, 1993.

[20] P.J. van Heerden, "Theory of optical information storage in solids," *Appl. Opt.*, 2(4):393–400, 1963.

[21] D. Gabor, "Associative holographic memories," *IBM Journal of Research and Development*, pages 156–169, 1969.

[22] D. Psaltis and F. Mok, "Holographic memories," *Sci. Am.*, 273(5):70–76, 1995.

[23] D. Gabor, "A new microscopic principle," *Nature*, 161:777, 1948.

[24] D. Gabor, "Microscopy by reconstructed wavefronts," *Proc. Roy. Soc*, A197:454, 1949.

[25] D. Gabor, "Microscopy by reconstructed wavefronts ii," *Proc. Phys. Soc*, B64:449, 1949.

[26] E. N. Leith, A. Kozma, J. Upatnieks, J. Marks, and N. Massey, "Holographic data storage in three-dimensional media," *Appl. Opt.*, 5(8):1303–1311, 1966.

[27] L. d' Auria, J. P. Huignard, C. Slezak, and E. Spitz, "Experimental holographic read-write memory using 3-D storage," *Appl. Opt.*, 13(4):808–818, 1974.

[28] D. L. Staebler, W. J. Burke, W. Phillips, and J. J. Amodei, "Multiple storage and erasure of fixed holograms in Fe-doped $LiNbO_3$," *Appl. Phys. Lett.*, 26(4):182–184, 1975.

[29] F. H. Mok, "Angle-multiplexed storage of 5000 holograms in lithium niobate," *Opt. Lett.*, 18(11):915–917, 1993.

[30] F. H. Mok, G. W. Burr, and D. Psaltis, "Angle and space multiplexed random access memory (HRAM)," *Optical Memory and Neural Networks*, 3(2):119–127, 1994.

[31] G. W. Burr, F. H. Mok, and D. Psaltis, "Angle and space multiplexed storage using the 90° geometry," *Opt. Commun.*, 117(1-2):49–55, 1995.

[32] D. Psaltis, "Parallel optical memories," *Byte*, 17(9):179, 1992.

[33] A. Pu and D. Psaltis, "High density recording in photopolymer–based holographic 3D disks," *Appl. Opt.*, 35(14):2389–2398, 1996.

[34] J. F. Heanue, M. C. Bashaw, and L. Hesselink, "Volume holographic storage and retrieval of digital data," *Science*, 265(5173):749–752, 1994.

[35] T. J. Hall, R. Jaura, L. M. Connors, and P. D. Foote, "The photorefractive effect – a review," *Progress in Quantum Electronics*, 10(2):77–145, 1985.

[36] D. L. Staebler, J. J. Amodei, and W. Philips, "Multiple storage of thick holograms in $LiNbO_3$," in *VII International Quantum Electronics Conference*, Montreal, 1972.

[37] S. Yin, H. Zhou, F. Zhao, M. Wen, Y. Zang, J. Zhang, and F. T. S. Yu, "Wavelength–multiplexed holographic storage in a sensitive photorefractive crystal using a visible–light tunable diode–laser," *Opt. Commun.*, 101(5-6):317–321, 1993.

[38] G. A. Rakuljic, V. Levya, and A. Yariv, "Optical data storage by using orthogonal wavelength–multiplexed volume holograms," *Opt. Lett.*, 17(20):1471–1473, 1992.

[39] C. Denz, G. Pauliat, and G. Roosen, "Volume hologram multiplexing using a deterministic phase encoding method," *Opt. Commun.*, 85:171–176, 1991.

[40] H. Lee, X.-G. Gu, and D. Psaltis, "Volume holographic interconnections with maximal capacity and minimal cross talk," *J. Appl. Phys.*, 65(6):2191–2194, 1989.

[41] K. Curtis, A. Pu, and D. Psaltis, "Method for holographic storage using peristrophic multiplexing," *Opt. Lett.*, 19(13):993–994, 1994.

[42] D. Psaltis, M. Levene, A. Pu, G. Barbastathis, and K. Curtis, "Holographic storage using shift multiplexing," *Opt. Lett.*, 20(7):782–784, 1995.

[43] G. Barbastathis, M. Levene, and D. Psaltis, "Shift multiplexing with spherical reference waves," *Appl. Opt.*, 35:2403–2417, 1996.

[44] H.-Y. S. Li and D. Psaltis, "Three dimensional holographic disks," *Appl. Opt.*, 33(17):3764–3774, 1994.

[45] G. W. Burr, X. An, F. H. Mok, and D. Psaltis, "Large–scale rapid–access holographic memory (Paper F2.3)," in *SPIE Optical Data Storage Topical Meeting*, San Diego, 1995.

[46] I. McMichael, W. Christian, D. Pletcher, T. Y. Chang, and J. Hong, "Compact holographic storage demonstrator with rapid access," *Appl. Opt.*, 35(14):2375–2379, 1996.

[47] K. Curtis and D. Psaltis, "Characterization of the Du-Pont photopolymer for 3-dimensional holographic storage," *Appl. Opt.*, 33(23):5396–5399, 1994.

[48] D. von der Linde and A. M. Glass, "Photorefractive effects for reversible holographic storage of information," *Appl. Phys.*, 8:85–100, 1975.

[49] F. S. Chen, J. T. LaMacchia, and D. B. Fraser, "Holographic storage in lithium niobate," *Appl. Phys. Lett.*, 15(7):223–225, 1968.

[50] J. B. Thaxter and M. Kestigian, "Unique properties of SBN and their use in a layered optical memory," *Appl. Opt.*, 13(4):913–924, 1974.

[51] K. Bløtekjaer, "Limitations on holographic storage capacity of photochromic and photorefractive media," *Appl. Opt.*, 18(1):57–67, 1979.

[52] G. C. Valley and M. B. Klein, "Optimal properties of photorefractive materials for optical data processing," *Opt. Eng.*, 22(6):704–711, 1983.

[53] S. M. Silence, R. J. Twieg, G. C. Bjorklund, and W. E. Moerner, "Quasinondestructive readout in a photorefractive polymer," *Phys. Rev. Let.*, 73(15):2047–2050, 1994.

[54] K. Meerholz, B. L. Volodin, B. Sandalphon Kippelen, and N. Peyghambarian, "A photorefractive polymer with high optical gain and diffraction efficiency near 100 percent," *Nature*, 371(6497):497–500, 1994.

[55] G. G. Malliaras, V. V. Krasnikov, H. J. Bolink, and G. Hadjiioannou, "Photorefractive polymer composite with net gain and subsecond response at 633 nm," *Appl. Phys. Lett.*, 65(3):262–264, 1994.

[56] J. W. Goodman, *Statistical Optics*, J. Wiley & Sons, 1985.

[57] H. Cramér, *Mathematical Methods of Statistics*, Princeton Univ. Press, 1946.

[58] S. Haykin, *Communication Systems*, J. Wiley & Sons, second edition, 1983.

[59] C. Gu, G. Sornat, and J. Hong, "Bit-error rate and statistics of complex amplitude noise in holographic data storage," *Opt. Lett.*, 21(14):1070–1072, 1996.

[60] C. Gu, F. Dai, and J. Hong, "Statistics of both optical and electrical noise in digital volume holographic data storage," *Electr. Let.*, 32(15):1400–1402, 1996.

[61] H. Hochstadt, *The functions of mathematical physics*, Dover, 1986.

[62] I. S. Gradshtein and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, fifth edition, 1994.

[63] E. T. Whitaker and G. N. Watson, *A Course of Modern Analysis*, Cambridge University Press, fourth edition, 1927.

[64] M. A. Neifeld and J. D. Hayes, "Error-correction schemes for volume optical memories," *Appl. Opt.*, 34(35):8183–8191, 1995.

[65] J. F. Heanue, K. Gurkan, and L. Hesselink, "Signal detection for page–access optical memories with intersymbol interference," *Appl. Opt.*, 35(14):2431–2438, 1996.

[66] M. A. Neifeld and W. C. Chou, "Information-theoretic limits to the capacity of volume holographic optical memory," *Appl. Opt.*, 36(2):514–517, 1997.

[67] G. W. Burr, J. Ashley, H. Coufal, R. K. Grygier, J. A. Hoffnagle, C. M. Jefferson, and B. Marcus, "Modulation coding for pixel–matched holographic data–storage," *Opt. Lett.*, 20(9):639–641, 1997.

[68] C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics*, Hermann & Wiley-Interscience, Paris, France, 1977.

[69] Hsin-Yu Sidney Li, *Holographic 3D Disks for optical data storage and artificial neural networks*, PhD thesis, California Institute of Technology, 1994.

[70] J. D. Jackson, *Classical electrodynamics*, J. Wiley & Sons, second edition, 1975.

[71] C. Gu, J. Hong, I. McMichael, R. Saxena, and F. Mok, "Cross-talk-limited storage capacity of volume holographic memory," *J. Opt. Soc. Am. A*, 9(11):1978–1983, 1992.

[72] K. Curtis, C. Gu, and D. Psaltis, "Cross–talk in wavelength–multiplexed holographic memories," *Opt. Lett.*, 18(12):1001–1003, 1993.

[73] A. Yariv, "Interpage and interpixel crosstalk in orthogonal (wavelength multiplexed) holograms," *Opt. Lett.*, 18(8):652–654, 1993.

[74] J. Trisnadi and S. Redfield, "Practical verification of hologram multiplexing without beam movement," *Photonic Neural Networks, Proc. SPIE*, 1773:362–371, 1992.

[75] K. Curtis and D. Psaltis, "Cross talk in phase-coded holographic memories," *J. Opt. Soc. Am. A*, 10(12):2547–2550, 1993.

[76] S. Campbell, X. M. Yi, and P. Yeh, "Hybrid sparse-wavelength angle multiplexed optical data storage system," *Opt. Lett.*, 19(24):2161–2163, 1994.

[77] D. Psaltis and A. Pu, "Holographic 3D disks," *The International Journal of Optoelectronics-Devices and Technologies*, 10(3), 1995.

[78] J. E. Ford, Y. Fainman, and S. H. Lee, "Array interconnection by phase-coded optical correlation," *Opt. Lett.*, 15(19):1088–1090, 1990.

[79] J. D. Kraus, *Antennas*, McGraw-Hill, 1950.

[80] K. Wagner and D. Psaltis, "Multilayer optical learning networks," *Appl. Opt.*, 26(23):5061–5076, 1987.

[81] L. Solymar and D. J. Cooke, *Volume holography and volume gratings*, Academic Press, 1991.

[82] H. C. Külich, "A new approach to read volume holograms at different wavelengths," *Opt. Commun.*, 64(5):407–411, 1987.

[83] H. C. Külich, "Reconstructing volume holograms without image field losses," *Appl. Opt.*, 30(20):2850–2857, 1991.

[84] J. W. Goodman, *Introduction to Fourier Optics*, Mc Graw–Hill, 1968.

[85] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley–Interscience, 1973.

[86] Geoffrey W. Burr, *Volume holographic storage using the* 90° *geometry*, PhD thesis, California Institute of Technology, 1996.

[87] Allen Pu, *Holographic 3D disks and optical correlators using photopolymer materials*, PhD thesis, California Institute of Technology, 1997.

[88] C. X.-G. Gu, *Optical neural networks using volume holograms*, PhD thesis, California Institute of Technology, 1990.

[89] K. Curtis and D. Psaltis, "Cross–talk for angle–multiplexed and wavelength–multiplexed image plane holograms," *Opt. Lett.*, 19(21):1774–1776, 1994.

[90] Pochi Yeh, *Introduction to photorefractive nonlinear optics*, Wiley & Sons, 1993.

[91] D. Psaltis, D. Brady, and K. Wagner, "Adaptive optical networks using photorefractive crystals," *Appl. Opt.*, 27(9):1752–1759, 1988.

[92] E. S. Maniloff and K. M. Johnson, "Maximized photorefractive data storage," *J. Appl. Phys.*, 70(9):4702–4707, 1991.

[93] A. Pu and D. Psaltis, "Shrinkage–insensitive holographic recording geometry," in *OSA Annual Meeting*, Rochester, NY, 1996.

[94] A. Pu and D. Psaltis, "Shift-multiplexed holographic 3–D disk system," in *International Symposium on optical memory and optical data storage*, Maui, Hawaii, 1996.

[95] J. J. Amodei and D. L. Staebler, "Holographic pattern fixing in electro-optic crystals," *Appl. Phys. Lett.*, 18(12):540–542, 1971.

[96] F. Micheron and G. Bismuth, "Electrical control of fixation and erasure of holographic patterns in ferroelectric materials," *Appl. Phys. Lett.*, 20(2):79–81, 1972.

[97] Y. Qiao, S. Orlov, D. Psaltis, and R. R. Neurgaonkar, "Electrical fixing of photorefractive holograms in $(Sr_{0.75}Ba_{0.25})Nb_2O_6$," *Opt. Lett.*, 18(12):1004–1006, 1993.

[98] D. Brady, K. Hsu, and D. Psaltis, "Periodically refreshed multiply exposed photorefractive holograms," *Opt. Let.*, 15(14):817–819, 1990.

[99] D. Psaltis, F. Mok, and H.Y.-S. Li, "Nonvolatile storage in photorefractive crystals," *Opt. Lett.*, 19(3):210–212, 1994.

[100] M. P. Petrov, S. I. Stepanov, and A. V. Khomenko, *Photorefractive crystals in coherent optical systems*, Springer–Verlag, 1991.

[101] M. A. Neifeld and M. McDonald, "Lens design issues impacting page access to volume optical media," *Opt. Commun.*, 120(1-2):8–14, 1995.

[102] M. A. Neifeld and M. McDonald, "Optical design for page access to volume optical media," *Appl. Opt.*, 35(14):2418–2430, 1996.

[103] A. Vander Lugt, "Packing density in holographic systems," *Appl. Opt.*, 14(5):1081–1087, 1975.

[104] X. M. Yi, S. Campbell, P. Yeh, and C. Gu, "Statistical analysis of cross–talk noise and storage capacity in volume holographic memory - image plane holograms," *Opt. Lett.*, 20(7):779–781, 1995.

[105] Kevin Curtis, *3–D photopolymer disks for correlation and data storage, and cross-talk in volume holographic memories*, PhD thesis, California Institute of Technology, 1994.

[106] X. M. Yi, P. Yeh, and C. Gu, "Statistical analysis of cross–talk noise and storage capacity in volume holographic memory," *Opt. Lett.*, 19(9):1580–1582, 1994.

[107] N. V. Kukhtarev, V. B. Markov, S. G. Odulov, M. S. Soskin, and V. L. Vinetskii, "Holographic storage in electrooptic crystals, I. Steady state," *Ferroelectrics*, 22:949–960, 1979.

[108] N. V. Kukhtarev, V. B. Markov, S. G. Odulov, M. S. Soskin, and V. L. Vinetskii, "Holographic storage in electrooptic crystals, II. Beam coupling - light amplification," *Ferroelectrics*, 22:961–964, 1979.

[109] A. C. Strasser, E. S. Maniloff, K. M. Johnson, and S. D. D. Goggin, "Procedure for recording multiple-exposure holograms with equal diffraction efficiency in photorefractive media," *Opt. Lett.*, 14(1):6–8, 1989.

[110] F. Mok, G. W. Burr, and D. Psaltis, "A system metric for holographic memory systems," *Opt. Lett.*, 21(12):896–898, 1996.

[111] G. Barbastathis and D. Psaltis, "Shift-multiplexed holographic memory using the two-lambda method," *Opt. Lett.*, 21(6):429–431, 1996.

[112] D. von der Linde, A. M. Glass, and K. F. Rodgers, "Multiphoton photorefractive processes for optical storage in $LiNbO_3$," *Appl. Phys. Lett.*, 25(3):155–157, 1974.

[113] F. Micheron and G. Bismuth, "Field and time thresholds for the electrical fixation of holograms recorded in $(Sr_{0.75}Ba_{0.25})Nb_2O_6$ crystals," *Appl. Phys. Lett.*, 23(2):71–72, 1973.

[114] S. Orlov, D. Psaltis, and R. R. Neurgaonkar, "Dynamic electronic compensation of fixed gratings in photorefractive media," *Appl. Phys. Lett.*, 63(18):2466–2468, 1993.

[115] S. Orlov, D. Psaltis, and R. R. Neurgaonkar, "Spatial and temporal characteristics of electrically fixed holograms in photorefractive strontium-barium niobate," *Appl. Phys. Lett.*, 64(7):824–826, 1994.

[116] G. Barbastathis, D. Psaltis, T. Chang, J. Hong, and R. R. Neurgaonkar, "Electrical fixing of angularly multiplexed holograms in SBN:75," in *OSA Annual Meeting '95*, Portland, Oregon, 1996.

[117] J. Ma, T. Chang, J. Hong, R. R. Neurgaonkar, G. Barbastathis, and D. Psaltis, "Electrical fixing of 1,000 holograms in SBN:75," in *Optical Data Storage '96*, Maui, Hawaii, 1996.

[118] J. Ma, T. Chang, J. Hong, R. R. Neurgaonkar, G. Barbastathis, and D. Psaltis, "Electrical fixing of 1,000 angle-multiplexed holograms in SBN:75," *Opt. Lett.*, 22(14):1116–1118, 1997.

[119] R. S. Cudney, J. Fousek, M. Zgonik, P. Günter, M. H. Garrett, and D. Rytz, "Photorefractive and domain gratings in barium titanate," *Appl. Phys. Lett.*, 63(25):3399–3401, 1993.

[120] M. Horowitz, A. Bekker, and B. Fischer, "Image and hologram fixing method with $(Sr_xBa_{1-x})Nb_2O_6$ crystals," *Opt. Lett.*, 18(22):1964–1966, 1993.

[121] Y. Qiao, D. Psaltis, C. Gu, J. Hong, P. Yeh, and R. R. Neurgaonkar, "Phase–locked sustainment of photorefractive holograms using phase conjugation," *J. Appl. Phys.*, 70(8):4646–4648, 1991.

[122] H. Sasaki, Y. Fainman, J. E. Ford, and S. H. Lee, "Dynamic photorefractive optical memory," *Opt. Let.*, 16(23):1874–1876, 1991.

[123] S. Boj, G. Pauliat, and G. Roosen, "Dynamic holographic memory showing readout, refreshing, and updating capabilities," *Opt. Let.*, 17(6):438–440, 1992.

[124] T. Dellwig, C. Denz, T. Rauch, and T. Tschudi, "Coherent refreshment and updating for dynamic photorefractive optical memories using phase conjugation," *Opt. Commun.*, 119:333–340, 1995.

[125] A. Yariv, "Three-dimensional pictorial transmission in optical fibers," *Appl. Phys. Lett.*, 28:88, 1976.

[126] R. W. Hellwarth, "Generation of time-reversed wave fronts by nonlinear refraction," *J. Opt. Soc. Am.*, 67(1):1–3, 1977.

[127] B. A. Fischer, editor, *Optical Phase Conjugation*, Academic Press, 1983.

[128] A. Yariv, *Optical Electronics*, Sounders College, 4th edition, 1991.

[129] J.-J. P. Drolet, E. Chuang, G. Barbastathis, and D. Psaltis, "Compact, integrated dynamic holographic memory with refreshed holograms," *Opt. Lett.*, 22(8):552–554, 1997.

[130] J.-J. P. Drolet, G. Barbastathis, and D. Psaltis, "Integrated optoelectronic interconnects using liquid-crystal-on silicon vlsi," in R. T. Chen and P. S. Guilfoyle, editors, *Optoelectronic interconnects and packaging*, volume CR-62, pages 106–131, 1996.

[131] D. A. Jared, R. Turner, and K. M. Johnson, "Electrically addressed spatial light modulator that uses a dynamic memory," *Opt. Lett.*, 16(22):1785–1787, 1991.

[132] D. A. Jared and K. M. Johnson, "Optically addressed thresholding very-large-scale-integration/liquid-crystal spatial light modulators," *Opt. Lett.*, 16(12):967–969, 1991.

[133] J. J. P. Drolet, J. S. Patel, K. G. Haritos, W. H. Xu, A. Scherer, and D. Psaltis, "Hybrid-aligned nematic liquid-crystal modulators fabricated on vlsi circuits," *Opt. Lett.*, 20(21):2222–2224, 1995.

[134] H.-Y. S. Li and D. Psaltis, "Alignement sensitivity of holographic 3-dimensional disks," *J. Opt. Soc. Am. A*, 12(9):1902–1912, 1995.

[135] Z. O. Feng and K. Sayano, "Compact read-only memory with lensless phase-conjugate holograms," *Opt. Lett.*, 21(16):1295–1297, 1996.

[136] Jean-Jacques P. Drolet, *Optoelectronic devices for information storage and processing*, PhD thesis, California Institute of Technology, 1997.

[137] D. P. Resler, D. S. Hobbs, R. C. Sharp, L. J. Friedman, and T. A. Dorschner, "High-efficiency liquid-crystal optical phased-array beam-steering," *Opt. Lett.*, 21(9):689–691, 1996.

[138] H. Kogelnik, "Coupled wave theory for thick hologram gratings," *Bell Syst. Tech. J.*, 48(9):2909–2947, 1969.

[139] T. W. Mossberg, "Time-domain frequency-selective optical data storage," *Opt. Lett.*, 7(2):77–79, 1982.

[140] U. P. Wild, S. E. Bucher, and F. A. Burkhalter, "Hole burning, Stark effect and data storage," *Appl. Opt.*, 24(10):1526–1530, 1985.

[141] X. A. Shen, E. Chaing, and R. Kachru, "Time-domain holographic image storage," *Opt. Lett.*, 19(16):1246–1248, 1994.

[142] M. Mitsunaga, N. Uesugi, H. Sasaki, and K. Karaki, "Holographic motion picture by Eu^{+3}:Y_2SiO_5," *Opt. Lett.*, 19(10):752–754, 1994.

[143] H. Lin, T. Wang, and T. W. Mossberg, "Demonstration of 8-Gbit/in^2 areal storage density based on swept-carrier frequency-selective optical memory," *Opt. Lett.*, 20(15):1658–1660, 1995.

[144] W. H. Hesselink and D. A. Wiersma, "Hahn- and stimulated photon-echo decay measurements in the lowest doublet-doublet electronic transition of triphenylmethyl in triphenylamine," *Chem. Phys. Lett.*, 50(1):51–56, 1977.

[145] H. C. Longuet-Higgins, "Holographic model of temporal recall," *Nature*, 217:104, 1968.

[146] D. Gabor, "Holographic model of temporal recall," *Nature*, 217:584, 1968.

[147] D. Gabor, "Improved holographic model of temporal recall," *Nature*, 217:1288–1289, 1968.

[148] F. Crick and C. Koch, "Are we aware of neural activity in primary visual cortex?," *Nature*, 375:121–123, 1995.

[149] F. Crick and C. Koch, "Consciousness and neuroscience," *Cerebral Cortex*, 1997, in press.

[150] R. Hecht-Nielsen, *Neurocomputing*, Addison-Wesley, 1990.

[151] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the theory of neural computation*, Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, 1991.

[152] C. M. Bishop, *Neural networks for pattern recognition*, Clarendon Press, 1995.

[153] S. H. Clearwater, B. A. Huberman, and T. T. Hogg, "Cooperative solution of constrained satisfaction problems," *Science*, 254:1181–1183, 1991.

[154] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, 4:219–227, 1985.

[155] A. Treisman, "Search, similarity, and integration of features between and within dimensions," *J. Exp. Psychol.: Hum. Perc. & Perf.*, 17:652–676, 1991.

[156] N. Kanwisher and J. Driver, "Objects, attributes, and visual attention: which, what, and where," *Current Dir. Psychol. Sci.*, 1:26–31, 1992.

[157] J. Braun and B. Julesz, "Dividing attention at little cost," *Percep. & Psychophys.*, 1997, in press.

[158] J. Nash, *Non-cooperative games*, PhD thesis, Princeton University, 1950.

[159] J. Nash, "Equilibrium points in n-person games," *Proc. Natl. Acad. Sci.* (USA), 36:48–49, 1950.

[160] D. Fudenberg and J. Tirole, *Game theory*, MIT Press, 1991.

[161] J. W. Weibull, *Evolutionary game theory*, MIT Press, 1995.

[162] H. B. Sarnat and M. G. Netsky, *Evolution of the nervous system*, Oxford University Press, second edition, 1981.

[163] J. Allman, T. Mac Laughlin, and A. Hakeem, "Brain-weight and life-span in primate species," *Proc. Natl. Acad. Sci.* (USA), 90:118–122, 1983.

[164] G. M. Shepherd, editor, *The synaptic organization of the brain*, Oxford University Press, third edition, 1990.

[165] D. H. Hubel, *Eye, brain, and vision*, Scientific American Library, 1988.

[166] S. M. Kosslyn, *Image and brain*, MIT Press, 1994.

[167] L. Itti, C. Koch, and E. Niebur, "Detecting salient objects in natural scenes using visual attention," in *Image Understanding Workshop IUW*, 1997, in press.

[168] S. Shimojo, O. Hikosaka, and S. Miyauchi, "Automatic and controlled attention detected by the line-motion effect," in T. Inui and T. J. McClelland, editors, *Attention and performance XVI*, Academic Press, 1997, in press.

[169] G. E. Hinton and J. A. Anderson, editors, *Parallel models of associative memory*, Lawrence Erlbaum Associates, 1981.

[170] J. L. Mc Clelland and D. E. Rumelhart, "A distributed model of human learning and memory," in D. E. Rumelhart and J. L. Mc Clelland, editors, *Parallel and distributed processing*, volume 2, pages 170–215, MIT Press, 1986.

[171] J. M. Guster, *Memory in the Cerebral Cortex*, MIT Press, 1995.

[172] V. B. Mountcastle, J. C. Lynch, A. Georgopoulos, H. Sakata, and C. Acuna, "Posterior parietal cortex of the monkey: command functions for operation within extrapersonal space," *J. Neurophysiol.*, 38:871–908, 1975.

[173] R. A. Andersen, "The role of the inferior parietal lobule in spatial perception and visual-motor integration," in F. Plum, V.B. Mountcastle, and S.R. Geiger, editors, *The Handbook of Physiology. Section 1: The Nervous System*, volume V. Higher Functions of the Brain Part 2, pages 483–518, American Physiological Society, Bethesda, MD, 1987.

[174] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Ann. Rev. Neurosci.*, 18:193–222, 1995.

[175] J. H. R. Maunsell, "The brain's visual world: representation of visual targets in cerebral cortex," *Science*, 270:764–769, 1995.

[176] S. Kinomura, J. Larsson, B. Gulyas, and P. E. Roland, "Activation by attention of the human reticular formation and thalamic intralaminar nuclei," *Science*, 271:512–515, 1996.

[177] J. M. Fuster, *The Prefrontal Cortex*, Raven Press: New York, 1989.

[178] M. K. Johnson and B. L. Chalfonte, "Binding complex memories: the role of reactivation and the hippocampus," in D. L. Schacter and E. Tulving, editors, *Memory Systems 1994*, pages 311–350, MIT Press, 1994.

[179] J. W. Rudy and R. J. Sutherland, "The memory-coherence problem, configural associations, and the hippocampal system," in D. L. Schacter and E. Tulving, editors, *Memory Systems 1994*, pages 119–146, MIT Press, 1994.

[180] M. L. Shapiro and D. S. Olton, "Hippocampal function and interference," in D. L. Schacter and E. Tulving, editors, *Memory Systems 1994*, pages 87–117, MIT Press, 1994.

[181] F. Crick, *The astonishing hypothesis*, Charles Scribner's Sons, New York, 1994.

[182] F. Crick and C. Koch, "Towards a neurobiological theory of consciousness," *Seminar in the Neurosciences*, 2:263–275, 1990.

[183] F. Crick and C. Koch, "The problem of consciousness," *Scientific American*, 267(3):153–159, 1992.

[184] R. Passingham, *The Frontal Lobes and Voluntary Action*, Oxford University Press, 1993.

[185] J. E. Bogen, "On the neurophysiology of consciousness, parts I and II," *Consciousness and cognition*, 4:52–62 & 137–158, 1995.

[186] W. Feller, *An Introduction to Probability Theory and Its Applications*, J. Wiley & Sons, third edition, 1968.

[187] G. L. Miller, S.-H. Teng, W. Thurston, and S. A. Vavasis, "Separators for sphere-packings and nearest neighbor graphs," *Journal of the ACM*, 44(1):1–29, 1997.