Minimum Information Holograms

Ravikanth Srinivasa Pappu

Bachelor of Engineering in Electronics and Communication Engineering,
Osmania University, 1991.
Master of Science in Electrical Engineering,
Villanova University, 1993.

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts & Sciences

at the

Massachusetts Institute of Technology

July 1995

Author:

Program in Media Arts and Sciences
July 19, 1995.

Certified by:

Stephen A. Benton
Allen Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Advisor

Accepted by:

Stephen A. Benton
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Minimum Information Holograms

Ravikanth Srinivasa Pappu

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on July 19, 1995 in partial fulfillment of the requirements for the degree of
Master of Science in Media Arts & Sciences

Abstract

The goal of this thesis is to characterize the redundancy inherent in a holographic fringe pattern and develop novel ways to eliminate it. The thesis consists of three independent but related parts.

A novel algorithm for computing compact holographic fringe patterns based on non-uniform sampling is presented. It is shown both analytically and via simulation that non-uniform sampling does indeed reduce the number of samples required to represent the holographic fringe pattern by at least 25% when compared to uniform sampling, with further gains that are dependent on the geometry of the object under consideration. The issues involved in determining the parameters are presented and discussed. Results demonstrating the performance of the algorithm are presented.

A simple model to determine the amount of information transferred to a beam of light by a spatial light modulator is presented. The model is based on information theoretic methods and the wave theory of light. Several examples and limiting cases are provided in order to demonstrate the validity of the model.

Finally, a comprehensive study of the pixel aspect ratios in holographic displays is presented. Canonical geometries and pixel aspect ratios are determined for both the spherical and the more general cylindrical case. In order to demonstrate the flexibility of the analysis, a holographic display that produces anastigmatic images using astigmatic horizontal and vertical holograms is presented.

Advisor: Stephen A. Benton

Title: Allen Professor of Media Arts and Sciences

This research was sponsored by the Television of Tomorrow Consortium at the MIT Media Laboratory.

Minimum Information Holograms

Ravikanth Srinivasa Pappu

The following people served as readers for this thesis:

Reader:

V. Michael Bove, Jr.
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Reader:

Neil Gershenfeld
Assistant Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Acknowledgments

“The time has come, the Walrus said, to talk of many things.....”

Finally, a page that is not in stilted academic prose like the rest of this thesis. Since many more people will read this page than the rest of the thesis, I guess I’ll be a tad more careful here.

I’m indebted to Steve Benton for taking the gamble and giving me a chance to work in the Spatial Imaging Group and for giving me the freedom to pursue a problem of my choice.

Thanks to my readers Mike Bove and Neil Gershenfeld for taking time to provide excellent feedback about various pieces of this thesis. Also, thanks to Neil for his terrific class “Modeling Nature” where many of my ideas about the marriage of holography and information theory were born.

To my friends in the Spatial Imaging Group, a big thank you! Thanks to Michele Henrion for teaching me holography in the summer of 1993, to John Sutter for being extremely helpful on both academic and social fronts in my first year, to Mike Klug and Mike Halle for being around to discuss half-baked ideas (my ideas, not theirs!), to Wendy “Babe 911” Plesniak for being an unofficial reader of my thesis and always being around to talk about anything under the sun, to John Underkoffler for being so eager to get to the bottom of any problem and asking me interesting and penetrating questions, to Melissa Yoon for being such a good friend, to Arno Klein for being such a crazy guy that it’s hard to be serious around him, and finally to Nobuhiro Kihara for being such a fun guy!

Thanks to my fellow electro-holographers — to Pierre St. Hilaire for teaching me everything about the display, to Mark Lucente for teaching me a lot about computational holography, and to Carlton “Anti-theory” Sparrell who always wanted to know if “it can be implemented in hardware”.

I acknowledge helpful discussions with Josh Smith, and also thank Rich Fletcher and Libby Shaw for helping me obtain the scanning probe microscope images on page 26.

A big thank you to my friends who made time away from the lab (?!!) fun: Raji, Sujal, Shami, Rama, Gaddam, Vani, Giri, Sripriya, Anu, Srikrishna, Kaushik, Suman,.....

Thanks to Linda Peterson and Santina Tonelli for keeping a watchful eye on the bureaucracy at MIT.

Finally, I want to thank my family who have constantly supported me in my endeavour to amass several rectangular pieces of paper with fancy writing on them. The next one will be the last.....perhaps.

Contents

1 Introduction

1 Introduction

1.1 Historical perspective

The problem of synthesizing free-standing spatial images has occupied the thoughts of researchers for over one and a half centuries. When optical holography was invented by Denis Gabor in 1949 [1], it was thought that a satisfactory solution to the problem of creating synthetic autostereoscopic images was finally found. However, holography has several limitations that keep it from being a viable medium for the routine presentation of 3D images. It seemed that the price that one had to pay for being able to see realistic three-dimensional images was to produce and view them under severely restricting conditions. Most of the restrictions arose from the use of coherent light to record and reconstruct the image.

In 1965, inspired by the work of Kozma and Kelly [2], Brown and Lohmann [3] realized that holograms could also be synthesized by computers by modelling the process by which optical holograms were recorded. A *computer-generated hologram* (CGH) may be defined as a numerical representation of a complex-valued interference pattern. Once the physics of holography was known, it was a simple matter to model and produce CGHs. Some of the earliest CGHs were amplitude-only masks, with applications in spatial filtering of images. The extension to computed display holograms is conceptually straightforward. However, simple calculations by Leith and his colleagues in 1965 showed that it would be extremely difficult, if not impossible, to compute display holograms using computational resources of the day. The single biggest hurdle was computing the enormous number of samples required to represent the hologram. Another obstacle was displaying the computed holograms in real-time. The possibility of producing CGHs that could be transmitted and reconstructed optically at a remote location was ruled out.

About six years ago, researchers at the MIT Media Laboratory Spatial Imaging Group designed and built the first system, dubbed the Mark I Holographic Video System, capable of displaying 3-D holographic images in real-time [4]. This advance in near-field computed display holograms was made possible by the adoption of information reduction strategies that pared down the raw

information content in the CGH to 2 megabytes from approximately 1 gigabyte. These strategies were: elimination of vertical parallax, reduction of horizontal and vertical display resolution, and reduction of image size and horizontal viewzone. Display of the reduced-information CGH was achieved by time-multiplexing the image of an acousto-optic modulator (AOM) to produce the final holographic image. The Mark I was capable of filling a volume the size of a golf ball with a horizontal-parallax only (HPO) three-dimensional holographic image. The image could be viewed in a 15° viewzone. Computation was carried out on a Connection Machine CM-2 parallel supercomputer. This work was reported in [5, 6]

A second-generation display, the Mark II, was built in 1993 [7, 8] and was based on the same principles as the Mark I. The important change was the introduction of parallelism into the optical architecture of the display. This allowed for a six-fold increase in the display capabilities of the system in terms of image volume and horizontal viewzone. The current configuration allows for an image which is 150mm wide, 75mm high, and 160mm deep. There are 144 HPO hololines each containing 256 kilosamples, giving a total of 36 megabytes per frame. The computing platform for this system is a custom digital video testbed called Cheops [9], that was developed at the MIT Media Laboratory. The main feature of Cheops that made it attractive for holographic video is its arbitrary, software-reconfigurable scan line length. All hologram computing research reported in this thesis was carried out on the Mark II using Cheops.

Much of the recent work in computed display holography has been done at the Media Laboratory during the course of the last three years by Lucente [10]. His recent doctoral dissertation concentrates on developing a new approach to hologram computing called diffraction-specific computation. This new approach has two driving forces - one, the rapid generation of holographic fringes and two, the production of novel holographic encoding techniques to enable bandwidth compression. In the interest of brevity, I will not describe his research here. The interested reader is referred to the dissertation for more details. Sutter [11] experimented with the computation of viewer-plane holograms for the Mark II.

1.2 Scope of the thesis

The goal of this thesis is to characterize the various elements of the information landscape in a holographic display. The primary application of such an effort would be in gaining insight into the redundancy inherent in computer-generated holograms and developing novel methods to eliminate it. Redundancy can be quantified in several ways and it is important to understand what these ways are and how the numbers they produce are related in order to develop new and efficient methods of generating and communicating holographic fringe patterns. Figure 1.1 depicts a typical holographic display as a communication channel. We will rely on Figure 1.1 on many occasions in the rest of the thesis. Various manifestations of the same holographic information are shown in the diagram.

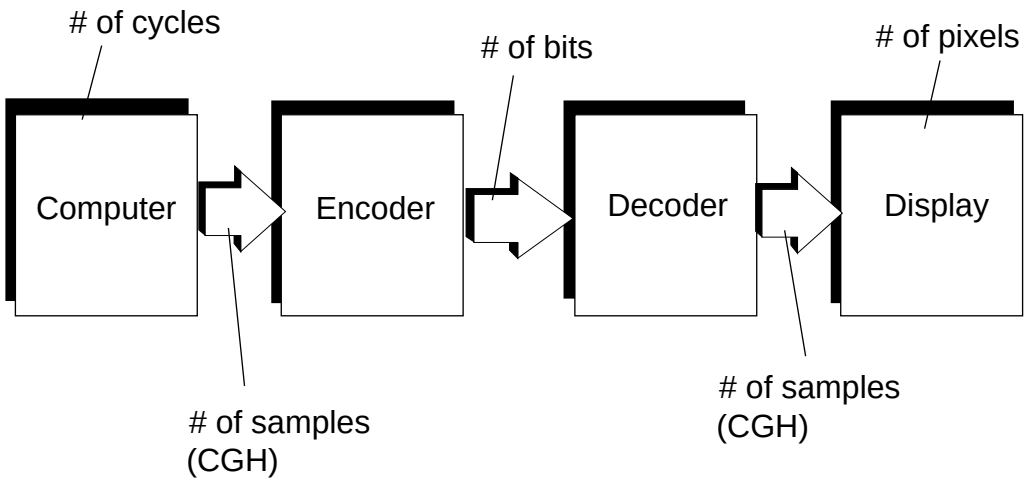


Figure 1.1 Different manifestations of holographic information in a typical display

The computer generates a CGH from a 3-D representation of the object which is then encoded and transmitted over the channel. The receiver decodes the signal and pumps the decoded data into the display for human consumption. The same holographic data is represented in very dissimilar forms at different points in the system, as is shown in the figure. It is convenient to think of the number of cycles required to compute the CGH as a form of data representation because we will treat it as a quantity requiring minimization. The relationships between the different forms of representation are not trivial and we will briefly discuss them here.

Hologram computing at interactive rates requires that the number of processor cycles used to compute each new hologram be as small as possible. Another quantity that requires minimization is the raw information that is represented by the number of samples in the computer-generated hologram. Currently used computing methods rely on uniform sampling to generate the CGH. This thesis explores minimization of the total number of samples using a nonuniform sampling technique.

The function of the encoder is to extract structure from the CGH and represent it in a compact and uniquely recoverable form. The bandwidth of the channel places a strong upper bound on the total number of bits that can be sent and recovered with arbitrarily small probabilities of error at the output. However, the relationship between the number of bits and the number of cycles is not intuitive. Shannon's channel coding theorem only demonstrates that there exist good codes that can be decoded with exponentially small probabilities of error, but provides no constructive method of generating them. Consequently, a code without some structure to it may take an infinite number of processor cycles to encode and decode. The decoded CGH is then presented to the beam of light via a modulator that is (usually) pixellated. Issues that require attention on the display end of the system are pixel sizes and dynamic range or bit-depth of each pixel.

In this thesis, we will focus attention on the following three issues:

- developing a nonuniform sampling technique to minimize the number of samples in the CGH — reducing the raw information
- looking at the process of information transfer in a spatial light modulator (SLM) — how to compute CGHs efficiently and
- characterizing pixel sizes corresponding to various reconstruction geometries.

The ultimate goal of this research is to completely characterize the information at every location in any holographic display. In order to build scaled-up, real-time, interactive holographic displays, it is essential to understand the complex interplay between distinct elements of the information space in the display. Admittedly, this thesis only begins that process and suggests avenues for deeper

exploration of the information landscape in a holographic display.

1.3 Overview

This document is organized as follows. The next chapter outlines the algorithm for nonuniform sampling and demonstrates, using simple algebra, that the total number of samples is indeed less than if the fringes were generated using uniform sampling techniques. Implementation of this algorithm is documented and results are provided to support the hypothesis. Chapter 3 is a discussion of an attempt to quantify the amount of information transferred to light in an SLM. The problem is approached both classically and quantum mechanically, with results being presented for the classical case. The format of information presentation to light in the display is the subject of Chapter 4. Pixel aspect-ratios are derived for both spherical and cylindrical optical systems and it is shown that cylindrical optical systems lead to a greater variety in pixel aspect-ratios. Chapter 5 summarizes the results of this thesis and restates the context in which the work was carried out. Future avenues for research are discussed. Figure 1.2 on the following page is a roadmap of this thesis and will be used often to pinpoint the reader's location in the information space of a holographic display.

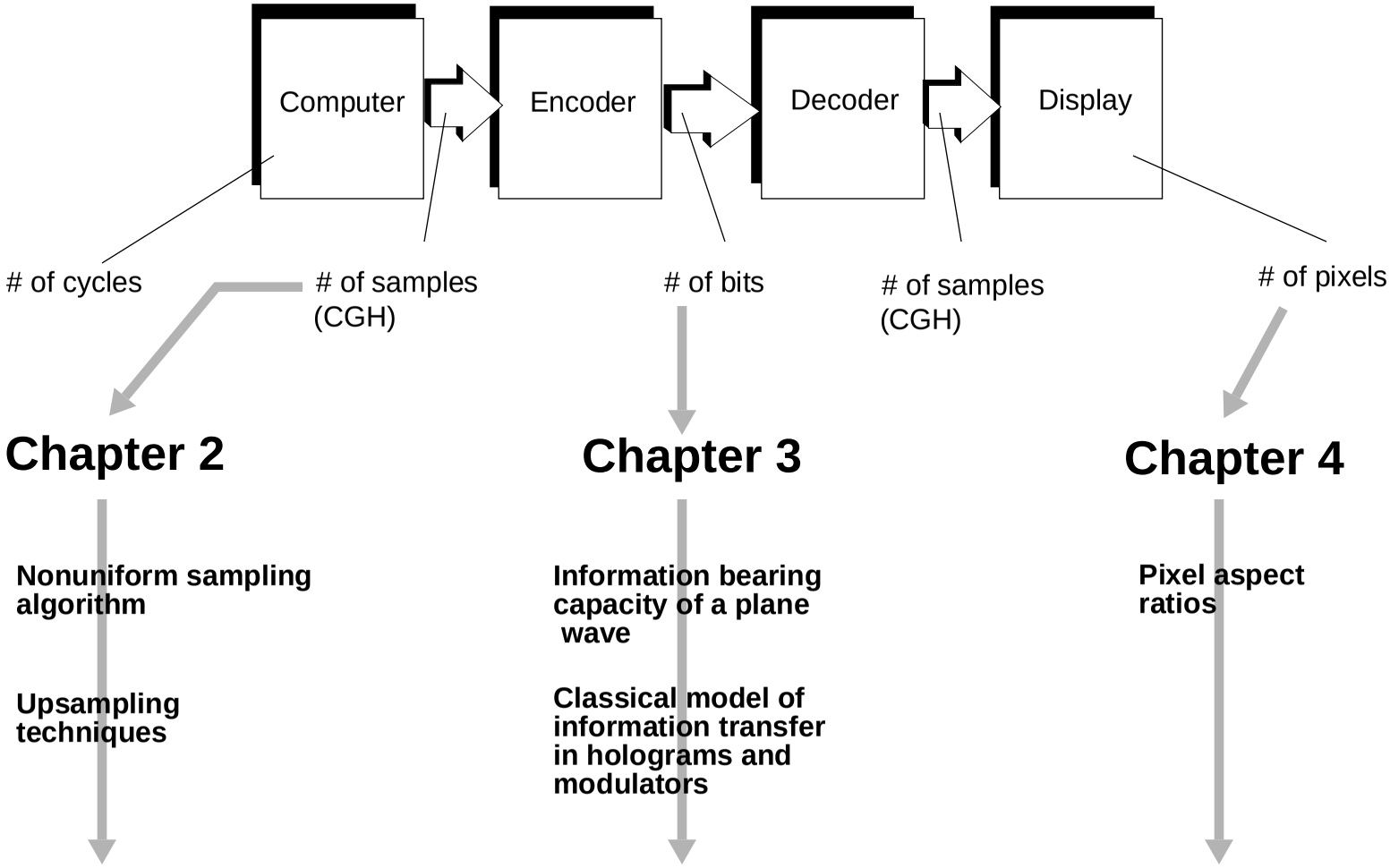
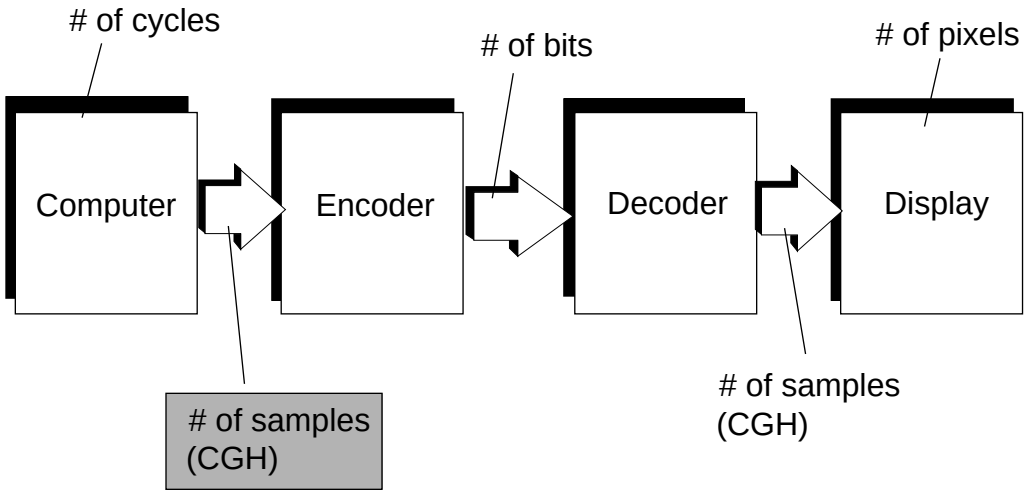


Figure 1.2: Thesis overview

2 Non-uniformly sampled computer generated holograms



2 Nonuniformly sampled computer generated holograms

2.1 Introduction

The idea of computing holographic fringe patterns and using them in optical setups is not new. It is almost as old as off-axis holography. This chapter is devoted to developing a different method of computing holographic fringe patterns. The technique presented in this section represents a departure from the conventional methods of thinking about CGHs. The primary difference between all previous methods and the method presented here lies in the way the hologram is sampled. The new method relies on a nonuniform sampling technique to represent different areas of the hologram plane.

A search of the literature has revealed no previous work in nonuniform sampling approaches of the type discussed in this chapter. There is, however, a paper by Vanderlugt [12] which mentions a related nonuniform technique of representing fringe patterns. The results of his paper will be presented in the next section, which presents previous work in computer-generated holography. The nonuniform sampling algorithm is developed in section 2.3. A simple algebraic proof that the algorithm does indeed reduce the number of samples necessary is also presented in this section. A visual corroboration of the algorithm is presented by means of scanning probe micrographs. Section 2.4 describes methods used to bring the subsampled CGH back to the original sampling rate and presents results for each method. The final section summarizes the material presented in this chapter and discusses approaches for future work.

2.2 Previous work

In this section, a concise review of the important techniques used to generate holographic fringe patterns in the past is presented. The list of techniques discussed serves two purposes. First, it demonstrates the historical progression of ideas related to representation of CGHs. Second, it shows that all but one of the methods rely on sampling the holographic fringe pattern uniformly. The one exception is [12] where Vanderlugt proves that the total number of samples required to represent a Fresnel (near-field) hologram is exactly the same as the number of samples needed for a Fourier (far-field) hologram provided the sample size is varied.

The problem of generating holographic fringe patterns may be stated very simply — how can the complex-valued fringe pattern be generated efficiently? In order to produce moving holographic images, the pattern has to be written onto an erasable and rewritable material which can be refreshed rapidly. All but one of the methods listed below rely on the physical model of interference of two waves — the object wave and the reference wave — to compute the hologram. A "recording geometry" is usually assumed and the complex amplitudes of the two waves are added in the hologram plane to produce the holographic fringe pattern. Then, a transparency has to be produced with the holographic fringe pattern on it. The encoding of the fringe pattern on a transparency is also dependent on the output devices available to print the pattern. It must be noted that it took a great amount of ingenuity to produce a high-resolution fringe pattern using fairly low-resolution output devices. We now proceed to review the different methods in roughly chronological order.

The binary detour-phase hologram: This type of computed hologram [3, 13] uses only two output levels in amplitude. The output plane is divided into $N \times N$ cells where each cell corresponds to a single Fourier coefficient of the Fourier Transform (FT) of the object. Each opaque cell has a rectangular aperture whose area is proportional to the amplitude of the Fourier coefficient. The phase is represented by translating the center of the opening to a different location within the cell as shown in Figure 2.1. The dimensions of the cell are $p \times q$ and the dimensions of the rectangular aperture are $a \times b$. The area ab is proportional to the magnitude of the transform and the displacement of the center of the aperture from the center of the cell is proportional to the phase of the transform. It may be shown that the diffracted far-field approximates the discrete Fourier transform of the computed complex amplitude under fairly general conditions. The diffracted far-field is equivalent to the required image. The binary masters are usually produced using a pen and ink plotter. However, this method doesn't make full use of the plotter resolution, since the number of plotter points in each cell must be large in order to keep quantization noise at bay.

Generalized binary detour-phase holograms: These holograms [14] are simply detour-phase holograms with a different method of amplitude and phase encoding. Instead of having a single aperture, as in the previous case, each cell is now subdivided into binary-valued subcell. This method allows for finer

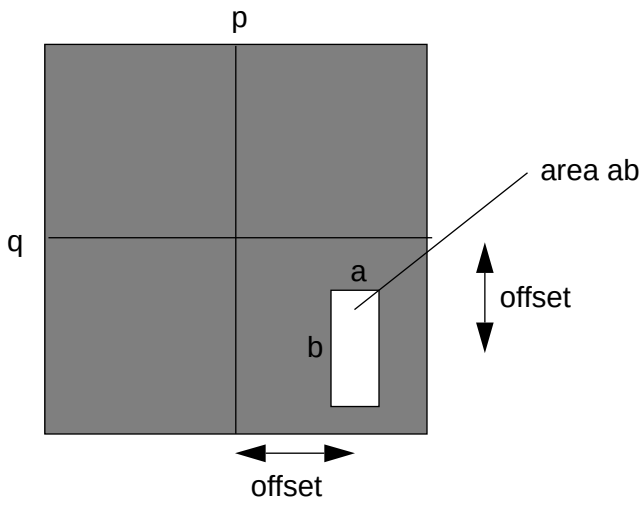


Figure 2.1: A single cell in a binary detour-phase hologram. The area of the aperture represents magnitude and the offsets represent phase of the Fourier transform

quantization of both amplitude and phase. This method is the holographic analog of halftoning and is illustrated in Figure 2.2 below.

Lee's method: This is an alternative method [15] of encoding amplitude and phase where the cell is divided into four vertical strips arranged side by side.

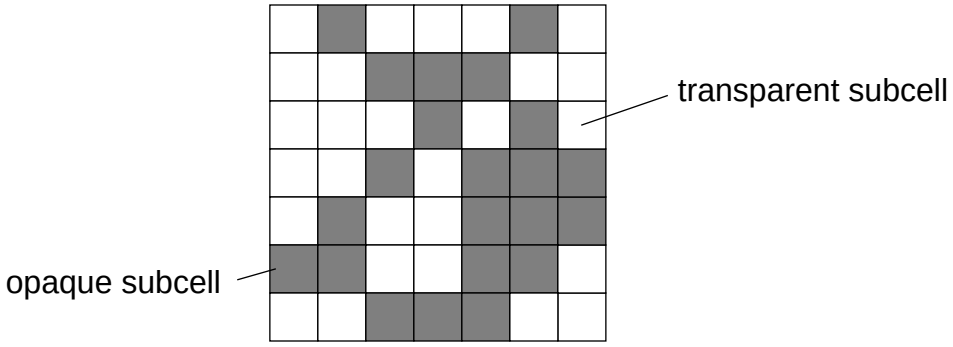


Figure 2.2: A single cell from a generalized binary detour-phase hologram

The four subcells contribute relative phases of 0, 90, 180, and 270 degrees as a consequence of their relative positions within the cell. The opacity of the subcells determines their amplitude contributions. In practice, it is possible to implement this method by keeping two out of four cells completely opaque, and adjusting the transmittances of the other two cells. Evidently, having two cells that are always zero implies that the encoding is redundant. Burckhardt [16] suggested that it was possible to implement this method using only three subcells, where each subcell had a different transmittance, possibly different from zero. Figure 2.3 (a) and (b) show typical cells from both methods. Note that the resultant amplitude and phase are identical in both cases but the configuration of subcells is distinct.

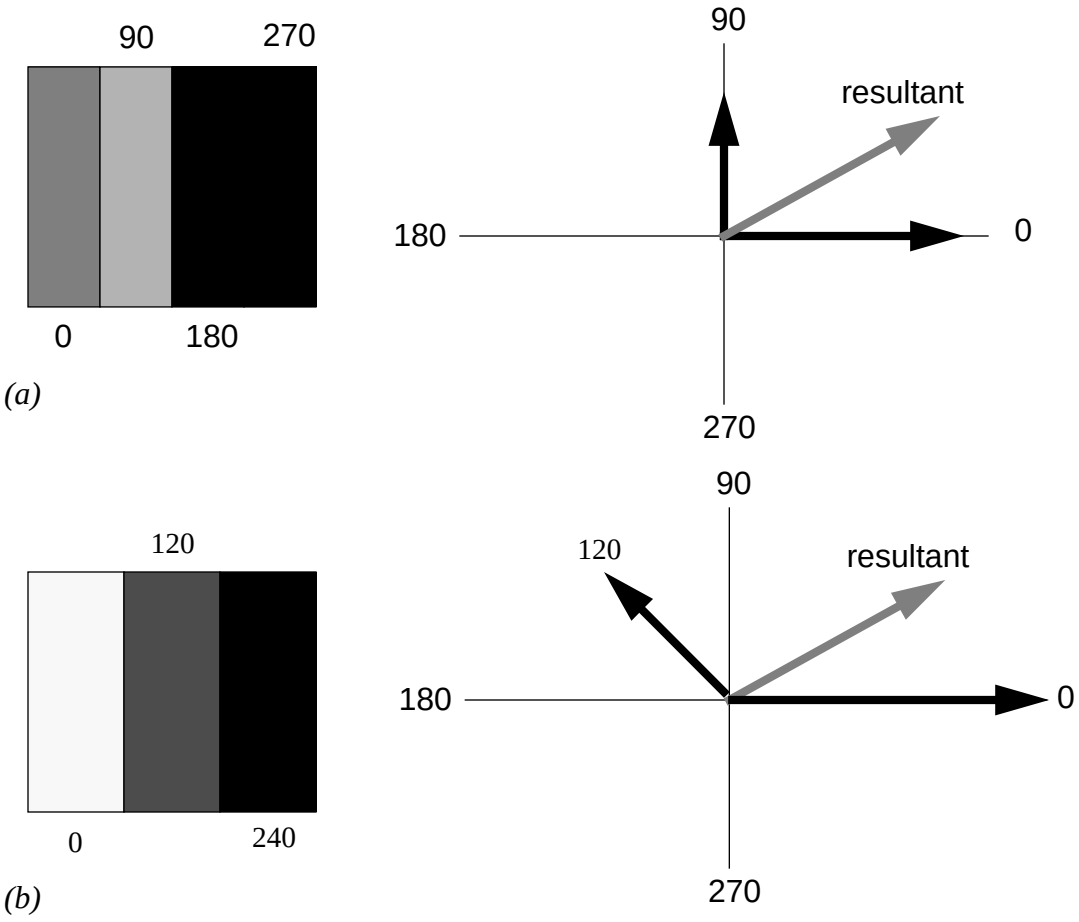


Figure 2.3: (a) Lee's method of encoding amplitude and phase (b) Burckhardt's modification. The numbers represent the relative phase shifts imparted by the various subcells.

The kinoform: When the object is diffusely illuminated the phases of the transform assume importance. This fact is exploited in the construction of the kinoform [17]. The CGH has all cells completely transparent so the moduli of the Fourier coefficients is arbitrarily set to unity. The phase of each cell in the kinoform is equal to the phase of the computed Fourier coefficient modulo 2π, which implies that the total phase variation over the area of the kinoform never exceeds 2π. Valuable dynamic range is conserved as a consequence of restricting the maximum phase excursion to be less than 2π. The kinoform is recorded on a photographic plotter which exposes a piece of film. The resulting master is photographed and bleached with a tanning bleach, which coverts any intensity variations into phase variations. If the processing is linear, then the bleached film contains a phase relief pattern which matches the phases of the Fourier coefficients.

The referenceless on-axis complex hologram (ROACH): This is an extension of the kinoform approach wherein both amplitude and phase are modulated [18]. Different layers of color film are used to encode both amplitude and phase. The magnitude of the Fourier transform is displayed on a cathode-ray tube and photographed using a filter of one color, say red. Then the phase of the transform is photographed using a blue-green filter. After processing, the image may be reconstructed by illuminating with red light. The red-sensitive layer modulates the amplitude of the illuminating light. The other layers pass red light but change the phase of the wavefront because their thickness is modulated with the phase of the Fourier transform. Thus, the emerging light is modulated by both the amplitude and the phase of the Fourier coefficients.

Several variations on the Fourier method have been explored in the last twenty years. Attempts have been made to compute the holographic fringe patterns of 3-D objects by slicing them in depth and computing and superposing the far-field hologram of each slice [19]. Iterative methods have been used to compute Fourier transform holograms [20]. The complexity of the objects used has kept pace with the increase in computing power available to the research community. However, Fourier transform holograms, by definition, are far-field holograms. In order to produce vivid 3D images, near-field — also known as the Fresnel regime — holograms have to be computed. We now turn our attention to efforts relating to the computation of near-field display holograms at the MIT Media

Laboratory.

Near-field holograms on a supercomputer: The first implementation of an algorithm to compute display holograms at the MIT Media Laboratory was performed on a Connection Machine - II supercomputer [5]. The approach was fairly straightforward in that it modelled the interference of light and computed the fringe pattern according to the model. Only HPO holograms were computed, which meant that the hologram was composed of *hololines* — different horizontal lines of the CGH — each of which diffracted light only in the left-right direction. The holograms were displayed on the Mark I MIT Holographic Video System.

Diffraction-specific computation: This is a novel approach by Lucente [10] that does away with computing the interference pattern prescribed by the physical model of interference. Instead, an attempt is made to compute the diffracting structure that will cause a given image to form. Diffraction-specific fringe computation is essentially a recipe for deciding what to put into a computed fringe pattern so that it achieves a particular goal. This goal is usually defined in terms of bending light through a certain angle or focusing it at some point, in the simple case. The CGH is treated as an array of small pieces, called *hogels*, each of which is the superposition of a certain number of basis fringes. The contribution of each basis fringe in a hogel is specified by the corresponding element of a vector called a *hogel-vector.* A vector-algebra representation of this process follows.

$$\underbrace{[w_1 \quad w_2 \ldots\ldots\ldots \quad w_N]}_{\text{hogel-vector}} \begin{bmatrix} basis_1 \\ basis_2 \\ \\ basis_N \end{bmatrix} = \underbrace{[fringe]}_{\text{hogel}}$$

The basis functions are derived by a process of simulated annealing. The elements of the hogel-vector are determined using geometrical optics. The hogel can then be generated by doing a simple vector-matrix multiply. This approach provides a significant improvement in the speed of computing because it does away with computing interference patterns. There is a one-time effort in computing basis fringes that are then stored in a lookup table. The physics of

diffraction is embedded in the basis fringes while the amount of light that goes in each direction is determined by the elements of the hogel-vector.

Optimal sampling of Fresnel Transforms: Finally, we present the one piece of work that proposes a flavor of nonuniform sampling. In [12], Vanderlugt states and proves that it is possible to represent a Fresnel transform — in other words, a near-field hologram — with exactly the same number of samples required to represent a far-field hologram *provided a specified nonuniform sampling method* was used. His method involves the use of unequal sample sizes (and hence, unequal sample spacing) to represent Fresnel transforms. It is interesting to observe that Vanderlugt anticipated one of the principal advantages of nonuniform sampling in almost the same words as the author of this thesis. A brief quote from [12] follows.

> "*This nonuniform sampling pattern is completely independent of the exact structure of the object and may be used in any application*"

Vanderlugt's work, however, was purely theoretical and was not implemented on any holographic display because there was only one real-time holographic display — The Mark I MIT Holographic Video system — in existence at the time of his writing the paper.

A quick review of the methods presented above reveals one feature common to all of them. The hologram area is populated with samples — either real or complex-valued — which are have a uniform spacing between them. In other words, they are all sampled at two times the maximum spatial frequency that occurs in the fringe pattern, following the prescription of the Shannon-Nyquist sampling theorem. This is perfectly justified if — and only if — every point on the hologram records every possible spatial frequency from zero to the maximum spatial frequency, say f_{max}. It is the author's view that such a situation rarely, if ever, arises. There are several areas of the hologram that record maximum spatial frequencies that are significantly less than f_{max}. In all the methods listed above, these areas would still be sampled at twice f_{max}, causing them to be oversampled. In this thesis, we explore a technique of sampling the hologram area nonuniformly and present some of the trade-offs involved. In the

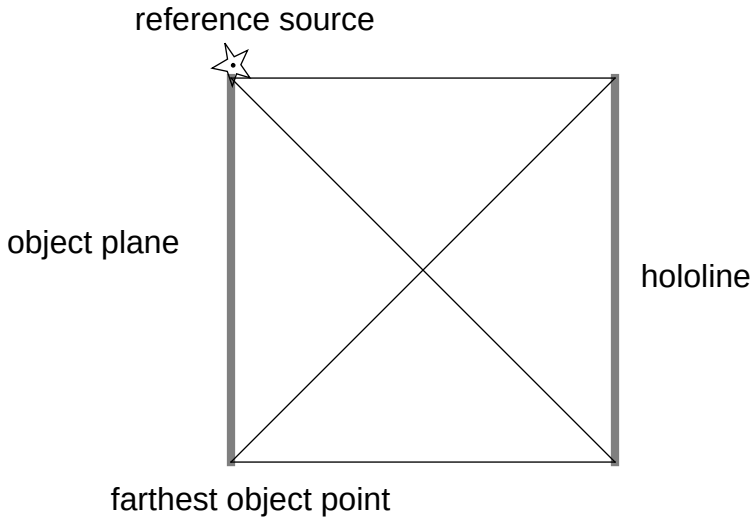


Figure 2.4: Maximum spatial frequency is approximately constant across the hololine.

next section, the proposed algorithm is developed and a proof that it does indeed reduce the total number of samples is presented.

2.3 Nonuniform sampling algorithm

In this section, we present the nonuniform sampling algorithm. Consider a typical hologram recording geometry as shown in Figure 2.4. The reference beam source is located at the very edge of the object plane — a technique possible in computational holography because the intermodulation (or "halo") term can explicitly excluded from the computation process. In the configuration shown, the maximum spatial frequency recorded on every point of the hologram is approximately the same. This is clear from the fact that the angles subtended by light from the farthest object point and reference beams at both ends of the hologram are the same, and this angle is approximately constant as the length of the plate is traversed. Constancy in spatial frequency is maintained under two conditions. First, the object should be two-dimensional. Second, the reference beam is diverging from a certain location. In the context of holovideo, both these conditions are not met. The objects are always three-dimensional and the reference beam has to be a plane wave because of the Bragg selectivity of the acousto-optic modulator. We conclude that the maximum spatial frequency

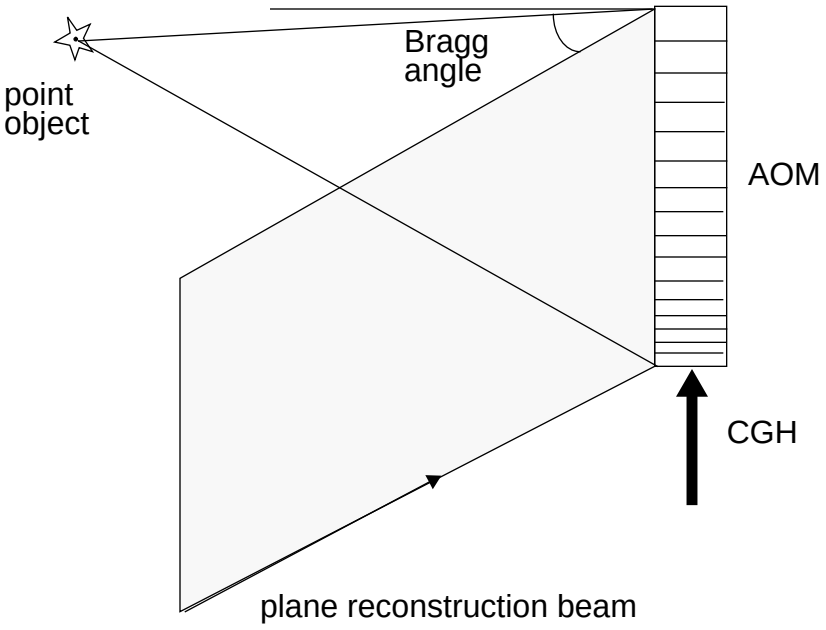


Figure 2.5: In holovideo, the reconstruction beam is plane and the spatial frequency across the length of the AOM is nonuniform.

recorded across the length of the hololine is not constant, but varies significantly from point to point. This is illustrated in Figure 2.5. We now proceed to develop the nonuniform sampling algorithm.

What follows is applicable to one hololine of an HPO hologram where the object is three-dimensional and lies behind the hologram plane.

Algorithm:

- Segment the hololine into N equal parts
- Identify the maximum spatial frequency f_{max}^{i} in each part ($i = 1, 2, \ldots\ldots, N$)
- Sample at $2f_{max}^{i}$ in each segment

Theorem:

The total number of samples in a hologram computed using the algorithm above

will always be less than or equal to an equivalent hologram computed using uniform sampling.

Proof:

Let the physical length of the hololine be P mm. The length of each segment is P/N mm. The total number of samples T_i in the ith segment is given by

$$T_i = (2Pf^i_{max})/N \text{ samples} \quad \text{2-1}$$

Since we picked the highest spatial frequencies in each segment, we may assume, without loss of generality, that

$$f^1_{max} \geq f^2_{max} \geq \ldots\ldots\ldots\ldots \geq f^N_{max} \quad \text{2-2}$$

The total number of samples $T_{nonuniform}$ in the hololine is given by

$$T_{nonuniform} = \sum_{i=1}^{N} T_i = \sum_{i=1}^{N} \left(\frac{2P}{N}\right) f^i_{max} \quad \text{2-3}$$

If the hologram had been sampled uniformly at $2f^1_{max}$, the maximum spatial frequency on the hololine, then the total number of samples in the line would be

$$T_{uniform} = 2Pf^1_{max} = \sum_{i=1}^{N} \left(\frac{2P}{N}\right) f^1_{max} \quad \text{2-4}$$

Using Eq. 2-2, it is clear that

$$T_{nonuniform} \leq T_{uniform} \quad \text{2-5}$$

This is an interesting result. It demonstrates that the total number of samples required to represent the nonuniformly sampled hologram is always less than or equal to the total number of samples required to represent an equivalent uniformly sampled hologram.

Visual justification:

In the previous subsection, we proved that the number of samples in a nonuniformly sampled CGH is smaller than the number of samples in a uniformly sampled CGH for the same object. The proof was based on the assumption that the maximum spatial frequency recorded is not constant across the length of the hololine. In order to see that this is indeed the case, the fringes on an optically recorded horizontal-parallax only hologram were observed using a scanning probe microscope. The hologram used was a 4 inch by 5 inch bleached rainbow hologram. Bleaching is necessary to convert intensity variation into a surface relief pattern. Three 100 micron by 100 micron scans were taken at three different locations along a horizontal line. These three scans are reproduced on the following page in Figures 2.6 (a), (b), and (c).

Perceptual justification:

The nonuniform sampling algorithm proposed may also be justified from a perceptual standpoint. In order to do this, consider the arrangement in Figure 2.7. A and B are two point sources of light located at different depths behind the hologram plane as shown. A plane reference beam is used to record the hologram. We will assume that the fringe patterns due to the two points do not overlap. It may be shown mathematically that the spatial frequencies recorded on the hologram due to point A are greater than those due to point B. In general, as the depth of the point increases behind the hologram plane, the spatial frequency components that the point contributes to the hologram decrease. Therefore, the maximum spatial frequency that the point contributes to the hologram may be used as a depth-dependent measure of information about the point. The nonuniform sampling algorithm described above is effectively performing the task of allocating samples on the hologram based on the depth of the points that contribute to it. It is, in some sense, giving priority to points that are closer to the viewer than those that are further away. Intuitively, one would expect that the viewer derives more three-dimensional information from objects that are closer than those that are far away. This intuition is supported by the fact that binocular disparity is zero for an object at infinite depth and greatest for an object that is at the near point of the viewer. The nonuniform sampling

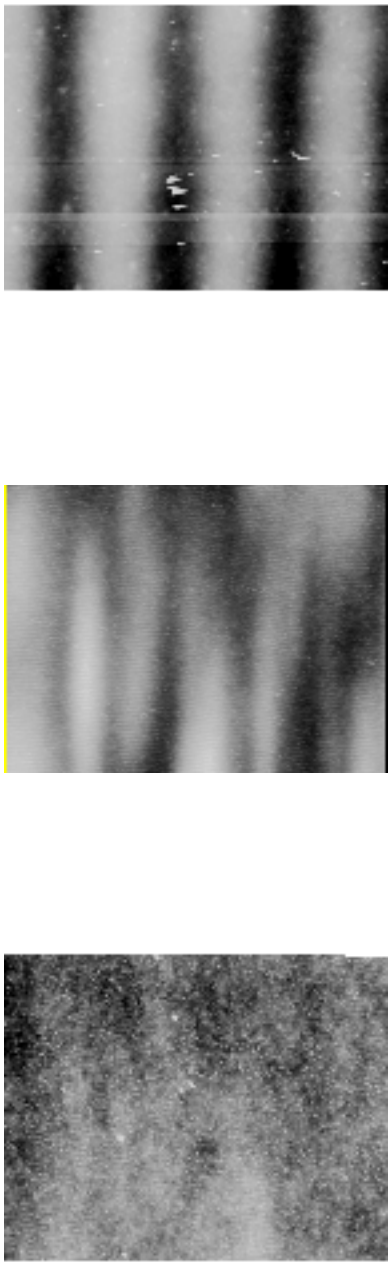


Figure 2.6: (a) The low spatial frequency portion of the hologram (b) the medium spatial frequency portion and (c) the highest spatial frequency portion. These three scans came from different locations on a single horizontal line in an HPO hologram. The horizontal extent of all these scans is about 7 microns.

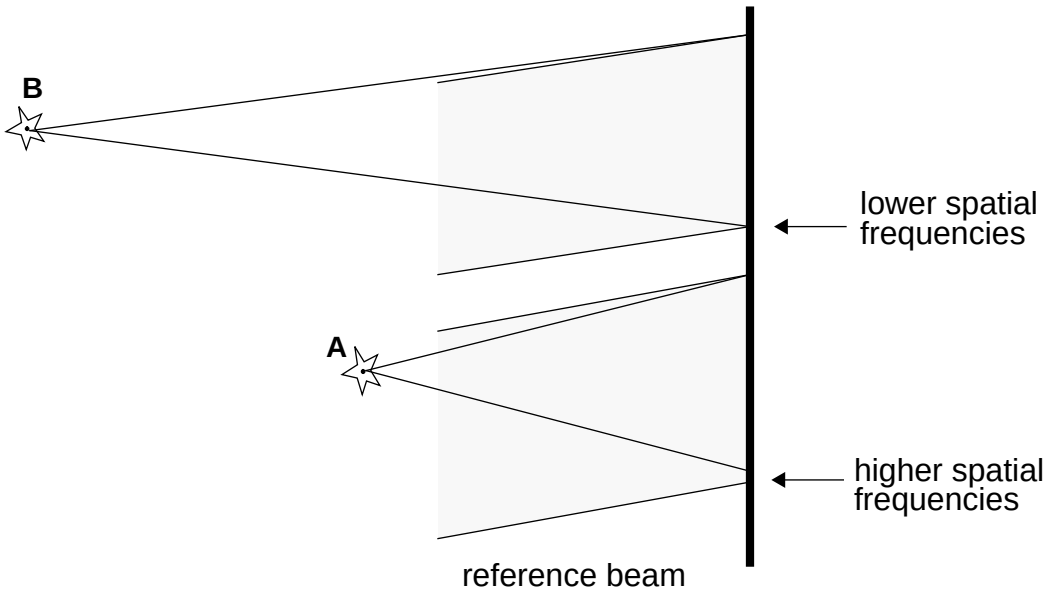


Figure 2.7: The maximum spatial frequency due to a point close to the hologram is higher than that due to a point further away.

algorithm achieves this by sampling different segments of the hologram at different rates.

2.4 Issues in nonuniform sampling

The implementation of any new algorithm brings with it a host of issues that must be understood and resolved in order to fully utilize the benefits that the algorithm provides. Nonuniform sampling is no exception. In this section, we present and discuss the most important issues that arise when nonuniform sampling is used to compute holograms for display applications. An example of an implementation is presented following which each of the issues is enumerated.

Example:

Once the object is available as a database of luminous points in space, the first priority is to determine the number of segments in the hololine. This is a non-trivial problem because the number of segments is object-dependent as

discussed below. However, experiments with various objects have shown that the savings begin to diminish when the number of segments exceed 128. This is evident from Figure 2.8 as well. In order to make the choice of the number of segments object independent, the number of segments was fixed at 128. After determining the number of segments, the maximum spatial frequency in each segment was determined using a geometrical optics approach. This process is very fast because it involves only a simple computation of the slope of the ray from an object point to the end of the segment. The maximum spatial frequency in each segment automatically determines the sampling rate and hence the pitch of the holographic fringe pattern. The hologram may then be computed in each segment using any method. The bipolar intensity method was used for the experiments in this thesis. The computed hologram is then stored in a format that will enable easy decompression. The format used for the examples in this thesis was quite simple. Each segment was accompanied by three pieces of information — the sampling rate in the segment, the number of samples in the segment, and the index of the segment. This is all that is necessary to upsample back to the maximum rate. The format is depicted in Figure 2.8.

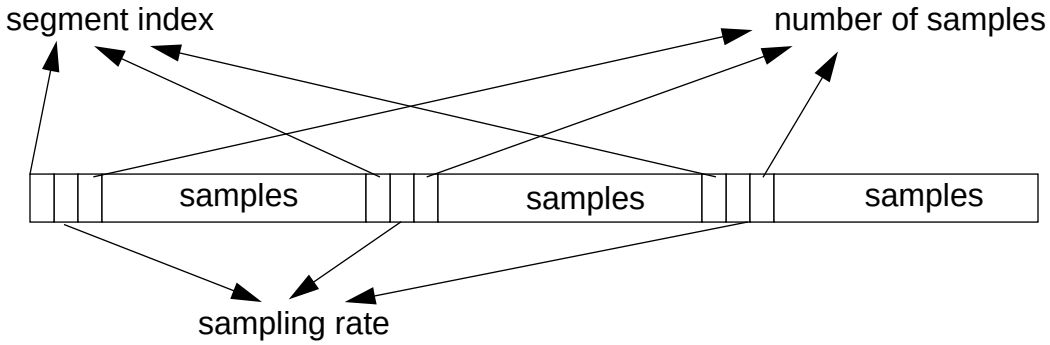


Figure 2.8: The file format of the computed hologram. The three pieces of header information enable upsampling to be performed on each segment.

Conceptual issues:

What is the optimal number of segments in a hololine? A look at the algorithm presented in the previous section indicates that a decision has to be made regarding the number of segments in any given hololine. Recalling Eq. 2.3, which is repeated here for convenience, we observe that the total number of

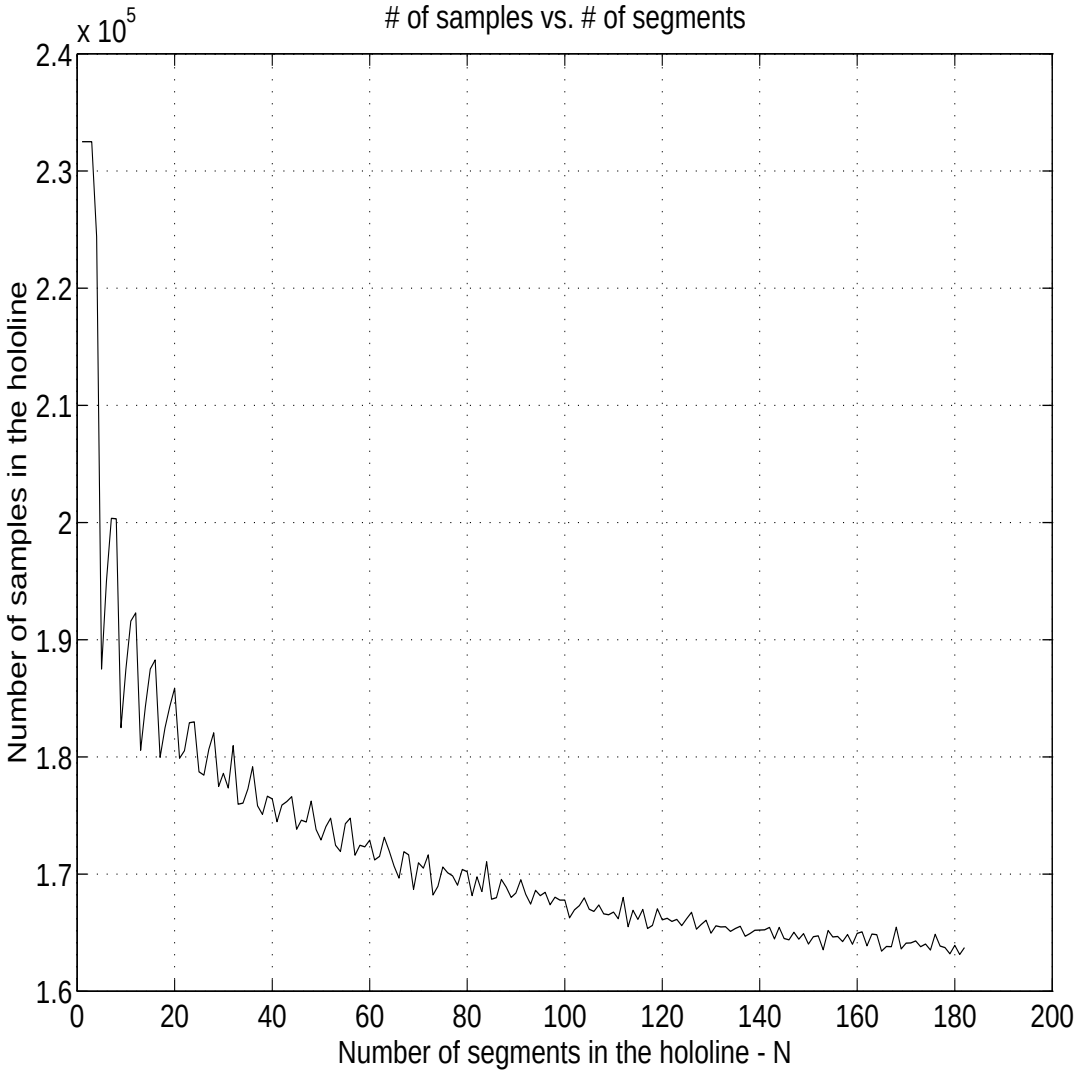


Figure 2.9: Variation in the number of samples in a hololine with the number of segments. The total number of samples reaches a steady state. The apparently periodic fluctuations on the left hand side are artifacts.

samples apparently decreases with increasing N.

$$T_{nonuniform} = \sum_{i=1}^{N} T_i = \sum_{i=1}^{N} \left(\frac{2P}{N}\right) f_{max}^{i}$$

What, then, prevents us from making the number of segments very large? There are three reasons for keeping N reasonably small. First, the maximum spatial frequency f^i_{max} in each segment depends on the size of the segment and the locations of its endpoints. f^i_{max}, therefore, depends on N in a non-trivial way. The relationship between the number of segments and the total number of samples in a hololine is examined graphically in Figure 2.9. Second, it must be kept in mind that the AOM requires that all the incoming data be sampled at a uniform rate. This is because the velocity of the acoustic wave in the AOM is constant regardless of sampling rate. In order to bring a low-rate segment back to a uniformly high rate, some form of interpolative filtering is necessary. A larger number of segments implies more time consumed in filtering. The interpolation issue is discussed more completely in the following section. Third, the law of diminishing returns takes effect as the number of segments is made larger. The savings in the number of samples diminishes and reaches a steady value when the number of segments crosses a certain threshold. Any number of segments above this threshold is more trouble than good. This is also illustrated in the example below.

Consider an object that is composed of a small number of points which are distributed randomly in the $x-z$ plane. Figure 2.9 is a plot of the total number of samples as the number of segments is varied from 1 to 180. The graph clearly indicates the drop in the number of samples as the number of segments is increased. Also, the total number of samples levels off when the number of segments becomes very large.

Interpolation: The output of the algorithm is a hololine which has N segments, each possibly sampled at a different rate. This hololine is fed into an acousto-optic modulator, which requires that its input signal be sampled at the same uniform rate. This is because the acoustic wave propagates down the crystal at a uniform velocity of 617 m/s regardless of the rate at which the signal was sampled. It is, therefore, necessary to convert the entire hololine to a uniform rate before sending it to the AOM. Conversion to a uniform rate can be achieved by interpolation, of which there are several kinds. Interpolation is simply the creation of some number of samples “in-between” the samples that one already has. No new information is created. Some of the kinds of interpolation are: low-pass interpolation, linear interpolation, and replication. In low-pass and linear

interpolation, the data to be interpolated is filtered by a specially designed digital filter to produce interpolated output [21]. Replication is a much simpler form of this where the number of extra samples required is produced by replicating existing sample values. For example, if a segment has 657 samples and it is required to have 1024 samples, then 367 samples randomly chosen in the data and replicated to produce a segment that contains 1024 samples. Replication is much faster than filtering if the operations are performed in software, but it also causes the shifting of samples from their original location—an artifact that does not accompany any other method of interpolation. Another method of interpolation is a technique borrowed from computer graphics known as the *Digital Differential Analyzer.* This algorithm uses a simple incremental algorithm to rapidly compute intermediate values given the endpoints of a line [22].

Object considerations: The number of points in the object, their density in the $x-z$ plane, and their depth from the hologram plane are all important factors in determining the total number of samples in any hololine. In this section, we look at a few limiting cases to get a feel for the behavior of the algorithm.

The worst case is when the object is a single line of points at the same depth with a very high density (number of points per unit length). Every segment will then receive a maximum spatial frequency contribution from some point in the line, which means that every segment will have to be sampled at a high rate. It makes sense in such cases to choose $N = 1$ so as to simplify the computation. It is indeed fortunate that not many of the objects of interest possess such a structure. On the other hand, objects with random point locations and depths, such as the one that generated Figure 2.9, make for very varied sampling rates leading to compact holograms. In this case, N should be chosen as large as possible to ensure greater savings. Of course, such objects also are not of much interest for display purposes either. Most of the objects that we will be interested in tread the middle ground between the two cases mentioned above. We, therefore, expect the savings in the number of samples to be somewhere between the two cases. There is also a slight cost in saving the header information as shown in Figure 2.8. The total number of bytes required to save the header information is $3N$ bytes, which is a very small number compared to the total number of samples.

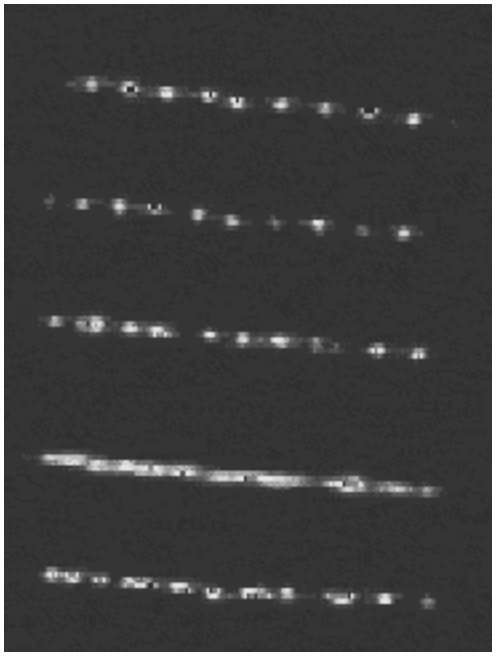
To summarize, the total number of samples decreases with the number of segments chosen, but the effort to interpolate goes as the number of segments increases. The structure of the object also determines what kind of sampling rates are necessary in each segment so it is important to look at the object and make a decision on the number of segments. It is only necessary to look at the gross structure of the object because the spatial frequency structure of the hologram depends solely on geometrical considerations. The exact structure is not important in this calculation. Once the sampling rates are determined, they can be stored for use again. In this sense, it is a one-time effort.

2.5 Results

The nonuniform sampling algorithm was implemented on an SGI Onyx workstation. Figure 2.10 shows the images of a single line of points which are spaced 1 mm apart in the x direction and 0.5 mm apart in the z direction. Because the points are uniformly spaced and span the whole object space, this is almost a worst-case scenario. There is always at least one object point that contributes the maximum possible spatial frequency in a particular segment. The hologram corresponding to the topmost image had a single segment and is equivalent to a fully computed hologram. The subsequent images had the number of segments indicated in the table alongside the figure. Two points must be noted. First, no interpolation was performed on the computed hologram. The hologram was displayed as computed, after being padded with zeros to make it appear like a 256 kilobyte hologram (as required by CHEOPS). The high tolerance of the AOM to changes in sampling rate is clearly evident. This feature indicates the possibility of doing less interpolative filtering than required. For example: if two consecutive segments have maximum spatial frequencies that differ by no more than 10%, it is possible to filter both of them with the same filter, or not at all as the case may be. Another point to note is that the images captured from the display have been through several format conversions before they ended up in Figure 2.10. The images look significantly better on the Mark II display than they do in Figure 2.10 for this reason.

2.6 Future directions

It was shown in the preceding sections that a nonuniform sampling algorithm may be used to reduce the number of samples in a computer-generated hologram by 25 to 30 percent depending on the object under consideration. This is the first time, to the best of the author’s knowledge, that nonuniformly sampled holograms have been used for display purposes. The approach closest



Segments	Samples
1	232500
64	229774
128	223280
256	218288
512	215736

Figure 2.10: Images of a single line of points produced by the algorithm. The first line consists of a single segment and is equivalent to a fully computed hologram. All subsequent lines are produced with varying number of segments (and hence samples) as indicated in the figure. All of these lines were produced without any interpolation, demonstrating the high tolerance of the AOM for varying sampling rates.

to the one in this chapter was proposed by Vanderlugt. It must be pointed out that the savings that accrue due to the nonuniform sampling algorithm are independent of any further savings that may be possible due to encoding schemes. The nonuniform sampling algorithm is lossless from an information theory standpoint. In this final section, we present two avenues for further study in the area of nonuniformly sampled computer-generated holograms.

Nonuniform segment widths: The algorithm described above used a certain number of segments *all of which were the same physical width, but sampled at different rates*. There is no reason why all the segments should be the same width. They could easily have been of different widths without increasing the

complexity of computation or interpolation. To see the implications of variable segment widths, let us reconsider Eq. 2-3. We replace the segment width P/N with w_i, the width of an individual chunk which may or may not be the same as w_j, for $i \neq j$. We then have

$$T_{nonuniform} = \sum_{i=1}^{N} T_i = \sum_{i=1}^{N} 2 w_i f_{max}^i \qquad 2\text{-}6$$

with the constraint that

$$\sum_{i=1}^{N} w_i = P$$

where P is the physical width of the hologram. This is an optimization problem. One possible approach is to decide the segment width adaptively based on the local spatial frequency gradient.

Bandpass sampling of CGHs: One assumption that was made in the formulation of the nonuniform sampling algorithm was that every segment has spatial frequency contributions all the way from zero to f_{max}^i. Studies have shown that this is not always true and that there are several segments where the spatial frequency range is quite small. Such a segment may be called a *bandpass* segment. There are well known techniques in communication theory that allow for sampling bandpass functions at a *smaller rate of twice the actual bandwidth* instead of at twice the maximum spatial frequency. One possible implementation of this technique is called heterodyned sampling. The bandpass signal is shifted in the frequency domain by multiplying by a sinusoid and then low-pass filtered to retain only the baseband. The baseband signal is then sampled at twice the bandwidth as prescribed by Nyquist. Some postprocessing is required to recover the actual samples from the baseband samples. The postprocessing involves representing the signal in the conventional analytic form and using a Hilbert Transform to recover the bandpass signal. This technique is documented in [23, 24]. The big gain lies in the significantly lower bandwidth communication between the host computer and the framebuffer made possible by using bandpass sampling.

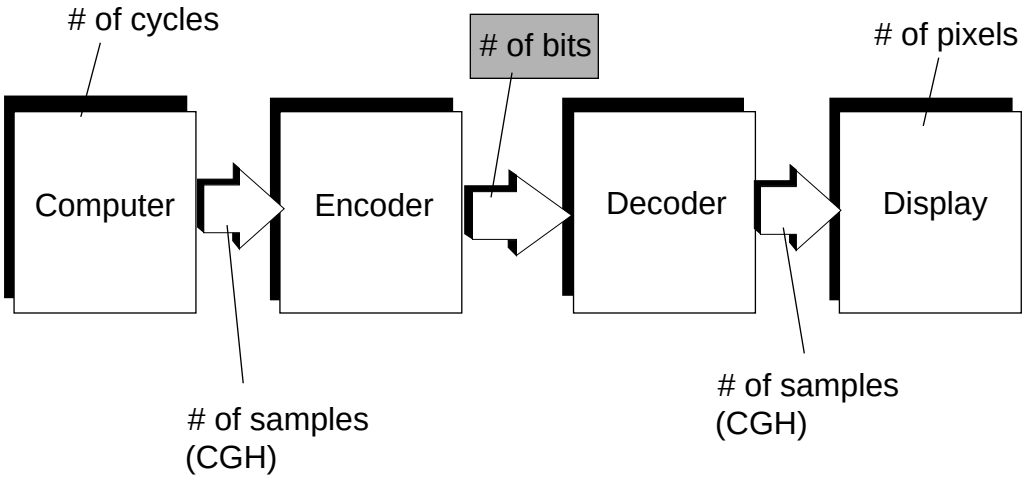
How is this technique of advantage to computational holographers? We recall

the arguments made in Section 1.2. One of the quantities that required minimization was the total number of samples in the CGH. Both nonuniform sampling and bandpass sampling achieve this goal — bandpass sampling is more effective. The number of samples that are sent to the framebuffer are reduced by using either method. If the framebuffer had no computational capability, then the entire exercise would be futile, because there would be no means of recovering a uniformly sampled stream of CGH data before it was input into the AOM. However, with an intelligent framebuffer like CHEOPS, it is possible to perform a great deal of signal processing on the data through the use of one or more specialized “daughter cards” — cards that are designed to implement a specific type of operation in hardware. Both nonuniform and bandpass sampling require the use of such hardware. Once again, it is emphasized that any advantages in computation and communication are independent of encoding methods that are used downstream in the communication chain.

Let us now consider how a nonuniform sampling scheme may be implemented if CHEOPS did not exist or, for some reason, was not available to us. The options available would be to use the radio-frequency (RF) portion of the holographic display to do some signal processing or to use optical means to decompress the data. An RF solution would involve changing the carrier frequency dynamically in order to match the sampling rate of each segment. This solution may be practically implemented. An optical solution would almost certainly involve more moving parts or a pulsed laser or both. Since one of the ultimate goals is to eliminate all moving parts from the display and make it inexpensive, both optical options do not seem to be very practical.

It is clear to the author that a great deal of work is still possible in simply reducing the number of samples in computer-generated holograms. This effort leads to CGHs that are lossless both perceptually and information theoretically. It is also necessary to study optical configurations that will allow the use of innovative algorithms in computational holography.

3 Modeling information transfer in SLMs



3 Modeling information transfer in spatial light modulators

3.1 Motivation

The history of holography -- both optical and computer-generated -- is replete with efforts to reduce the amount of "information" required to produce a "visually or perceptually lossless" image. The methods of attack have been very varied but the goal has always been the same. These methods usually fall into one of two categories - optical or computational. Optical methods use novel optical configurations to reduce the amount of information in fairly obvious ways. Computational methods introduce a greater measure of subtlety into the process of information reduction by using information theory and signal processing.

The big leap in reduction of information in a hologram came with the introduction of horizontal-parallax only (HPO) imaging. Much has been said about HPO holograms [25, 26] and the reader is referred to the literature for more details. At about the same time, Lin [27] proposed that smaller, identical holograms could be tiled to produce a larger hologram. The reconstructed image exhibited a significant loss of resolution and was therefore not perceptually lossless. In his concluding remarks, Lin noted that a "*practical method of electronic processing of even a very small hologram*" was still to be found. Though the problem of computing holographic fringe patterns has been solved, important questions about the nature of holographic information remain to be answered. Several other researchers developed similar schemes to reduce the amount of information in a hologram.

The bag of computational tricks used to reduce the information content in a hologram contains several methods. Among them are: using the Fast Fourier Transform to compute far-field holograms and manipulating phase to reduce speckle artifacts in computed holograms. More recently, Lucente [12] worked on developing encoding schemes specific to holographic data. All of the schemes outlined above approached the problem by asking the question: *A hologram contains a certain amount of information. What is the best way to reduce it*? The author of this thesis believes that the key to efficient CGH computation lies in approaching the problem from the bottom up — *to determine the minimum amount of information required to produce a desired*

visual sensation. The key to answering this question lies in understanding the mechanism of information transfer from a CGH to a beam of light. This chapter is devoted to a preliminary exploration of this important topic.

The structure of this chapter is as follows. The next section presents a brief introduction the principles of information theory. In section 3.3, the information bearing capacity if a plane wave is determined. The information bearing capacity of a plane wave is only of tangential interest to us because we are not engaged in the production or communication of plane waves. In a sense, we are end users of plane waves that have already been produced in a laser. However, determining the information bearing capacity of a plane wave provides some insight into information theoretic methods. Section 3.4 treats the light as a collection of plane waves and assumes that the SLM transfers information to the plane waves to produce a different set of plane waves. The amount of information required to achieve this input-output transformation is determined. The premise of this chapter is that if the minimum amount of information is determined, then the next step would be to isolate a class of waveforms that can deliver this amount of information to a beam of light efficiently.

3.2 Information theory principles

Information theory [28, 29, 30] is the science that is concerned with two fundamental questions — what is the ultimate data compression possible and what is the ultimate transmission rate of a a communication channel? Information theory deals with estimating the information content of mathematical objects, which in our case, are computer-generated holograms. By its very nature, information theory usually never provides constructive methods of achieving the limits it prescribes. Achieving the limits is left entirely to the ingenuity of researchers and their insight into the mathematical object at hand. The most often used tools in information theory are the concepts of *entropy*, *relative entropy*, and *mutual information*. They will be defined here for use later in the chapter.

The *entropy* of a random variable X with probability density function $p(x)$ is defined by

$$H(X) = -\sum p(x) \log_2 p(x) \qquad 3\text{-}1$$

By definition, the entropy is a positive quantity measured in units of *bits*. It may be interpreted as the average number of bits required to describe the random variable X. It is also a measure of the average uncertainty of the random variable. One of the primary results in information theory states that codewords used to describe mathematical structures must have an average length that is greater than or equal to the entropy of those structures. The most efficient coding schemes will generate codewords whose average length is equal to the

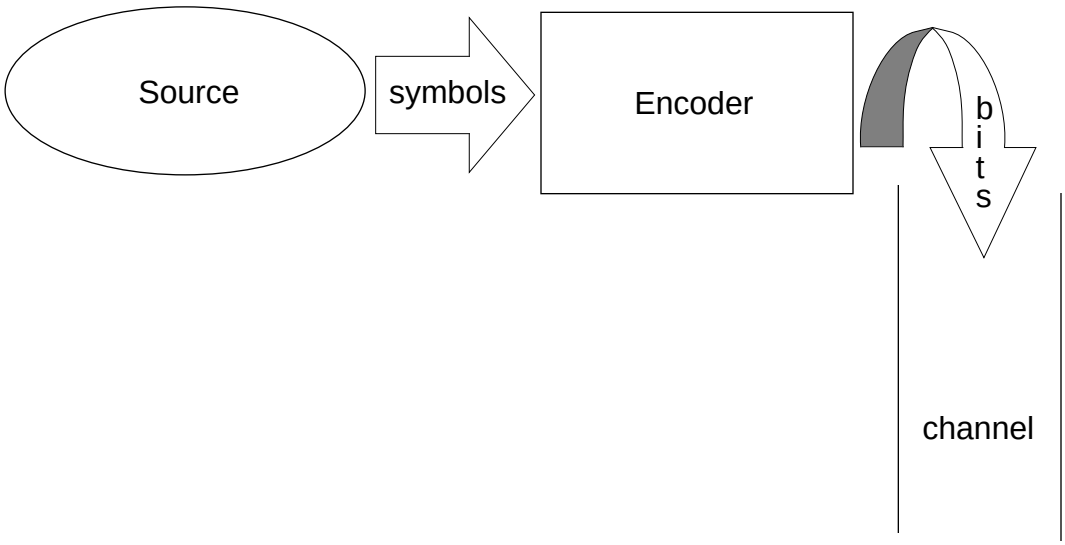


Figure 3-1: The encoder acts as an information "impedance-matching" device between the source and the channel

entropy of the source. The encoder may be interpreted to be an information "impedance-matching" device in that it matches the entropy of the source to the capacity of the channel. This is shown in Figure 3.1. In the context of holovideo, the channel capacity is determined by the output bandwidth of the frame buffer. In order to have a metric for comparing different encoding schemes, we need to know what the entropy of the source (the CGH) is. There are two approaches that can be used to determine this quantity. It is possible to look at a large number of holograms statistically and determine their probability density function empirically and use it to estimate the entropy. The other method, which is more elegant, is to determine the entropy of the hologram from first

principles. The latter approach is taken in the following sections.

Relative entropy is the distance between two probability density functions. It is a measure of the of the inefficiency of using a distribution q to describe something when the true distribution is p. It is also known as the *Kullback-Liebler* (KL) distance in the statistical literature. It is defined as

$$D(p \| q) = \sum_{allx} p(x) \log_2 \left(\frac{p(x)}{q(x)} \right) \qquad 3\text{-}2$$

For example, if we knew the true density function of the random variable, we could construct a code with average codeword length $H(p)$ bits. If, instead, we used a code for a distribution q, we would need $H(p) + D(p \| q)$ bits on average to describe the random variable. The relative entropy will be used later to determine the amount of information transferred in a hologram or spatial light modulator.

The *mutual information* is another quantity that is used often in describing the interplay between two variables. Mutual information is related to conditional probability. The uncertainty of an event occurring decreases if we know that another event has already occurred. The definition of mutual information is a quantitative statement of this fact.

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)\,p(y)} \qquad 3\text{-}3$$

$H(X|Y)$ is the conditional entropy which may be computed using Eq. 3.1 Since $H(\ )$ is always positive, it is clear that prior knowledge reduces entropy — a fact that makes intuitive sense. Having defined entropy and mutual information as they will be used in this chapter, we proceed to look at the classical approach to determining the entropy of a hologram.

3.3 Information bearing capacity of a plane wave

In this section, the information bearing capacity of a plane wave is determined. This material in this section is a digression from the actual goal of this chapter because we are not called upon to generate plane waves synthetically. We simply use the plane waves that are generated by a laser. The notation to be used

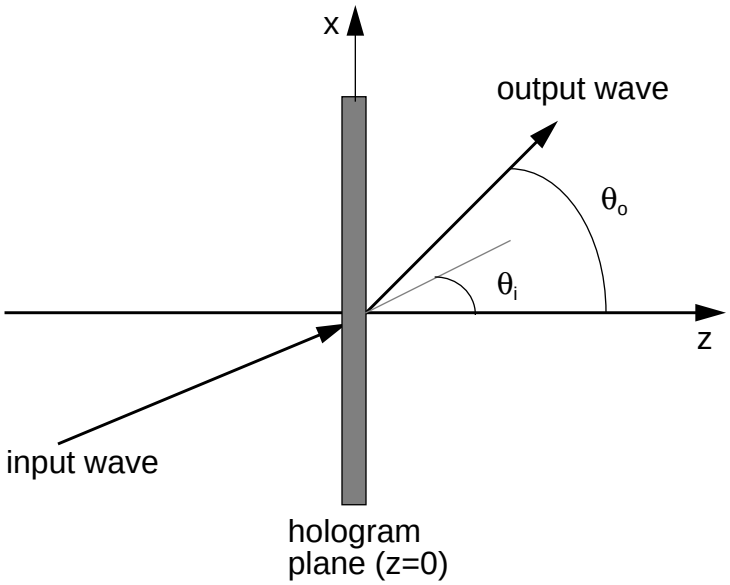


Figure 3.2: Convention for angles

in this section is as follows. The hologram is always assumed to be at $z = 0$ and all angles are measured with respect to the positive direction of the z-axis as shown in Figure 3.2. The input beam is assumed to come in from the left and leave the hologram plane travelling to the right. We will work only with one-dimensional plane waves with HPO holograms in mind.

The one-dimensional plane wave representation that will be adopted is

$$U(x) = a_i e^{j2\pi f_i x} \qquad 3\text{-}4$$

a_i is the *amplitude* of the plane wave and f_i is the *spatial frequency* of the input plane wave. The spatial frequency is related to the direction of propagation of the wave via Eq. 3-5 below.

$$f_i = (\sin\theta_i)/\lambda \qquad 3\text{-}5$$

where λ is the wavelength of light being used.

Let us denote the entropy of the input wave by H_i. We require a method by which we can describe the input entropy in terms of the physical parameters of the wave, such as its spatial frequency or amplitude. How much information does a plane wave contain? Intuitively, one would feel that a single plane wave requires an infinite amount of information to specify both the wavelength and direction exactly. However, ideal plane waves do not exist. They are a mathematical idealization. There is always a spatial frequency uncertainty (which is intimately related to the angular uncertainty, as shown below) associated with a wave that depends on the size of the source that produced the wave. Let us denote this quantity by Δf_i for the input wave. The relationship between f_i and θ_i is

$$f_i = \frac{\sin(\theta_i)}{\lambda} \tag{3-6}$$

Differentiating Eq. 3-6 gives us the relationship between the spatial frequency uncertainty Δf_i and the angular uncertainty $\Delta\theta_i$.

$$\Delta f_i = \frac{\cos(\theta_i)}{\lambda}\Delta\theta_i \tag{3-7}$$

Keeping in mind that a plane wave with a lower Δf_i requires greater precision to specify it, we define a measure of information of a plane wave as the logarithm to the base two of the ratio of f_i to Δf_i. This is simply the total number of bits in the binary representation of the ratio. Clearly, this ratio is inversely proportional to the spatial frequency — and therefore, angular — resolution. The ratio may also be interpreted as a “quality factor” for the plane wave. The higher the quality factor, the greater the number of bits required to describe the plane wave. The entropy, using Eqs. 3-12 and 3-13, is:

$$H_i = \left(\frac{f_i}{\Delta f_i}\right)\log_2\left(\frac{f_i}{\Delta f_i}\right) = \left(\frac{\tan\theta_i}{\Delta\theta_i}\right)\log_2\left(\frac{\tan\theta_i}{\Delta\theta_i}\right) \text{ bits} \tag{3-8}$$

The entropy, as defined by Eq. 3-8, depends on both the spatial frequency (or, equivalently, angle of propagation) and the spread in the spatial frequency. A limiting case of interest is: $\Delta f_i \to 0$ implies that $H_i \to \infty$. This is related to infinite spatial frequency resolution The behavior of the entropy is plotted against spatial frequency resolution and input angle in Figure 3.3. This analysis also highlights that fact that an origin and a coordinate system are always required when information is measured.

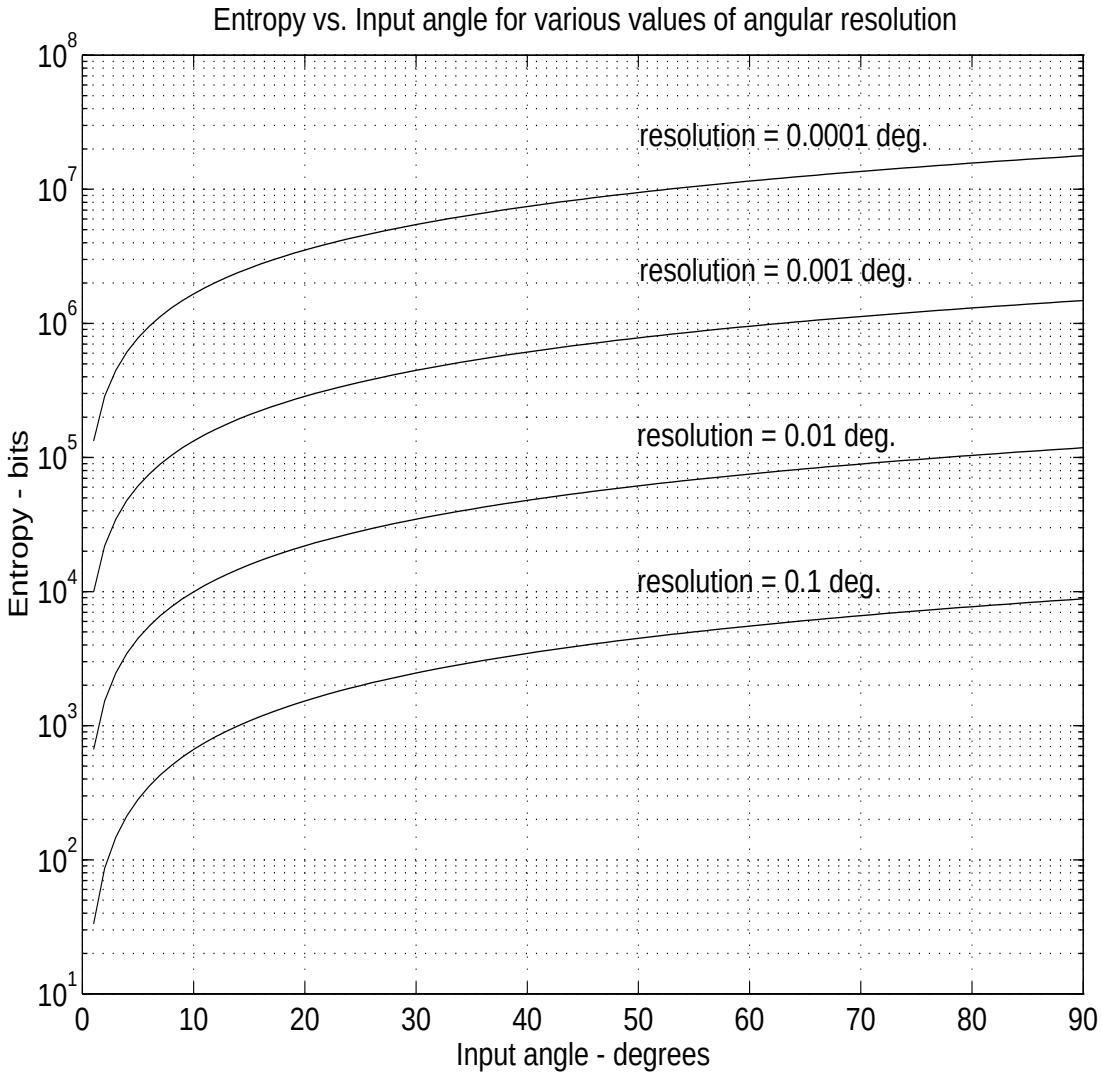


Figure 3.3: The behavior of entropy as the input angle is varied for various angular resolution values.

3.4 Classical theory of information transfer

We now focus attention on the problem of determining the information-theoretic lower bound on the amount of information in a holographic fringe pattern. There are three reasons why such an endeavour should be undertaken. First, such a measure provides a clear way of comparing different encoding schemes. Second, it is expected that such an analysis will provide some insight into the dynamic range requirements on holographic fringe patterns. Finally, it may also provide insight into designing the encoding scheme itself — although this is a less practical expectation. The approach to determining the information content of a fringe pattern is straightforward. The "information distance" between the input and output plane wave distributions will be determined by invoking a variant of the Kullback-Liebler (KL) distance defined in Eq. 3-2. When this distance is measured in bits, it represents the amount of information added to the input distribution of plane waves by the grating. It should also be remembered that the KL distance is a lower bound on the information content of a holographic fringe pattern. Determining KL distance does not predicate the existence of an encoding scheme that enables us to reach the lower information bound. The assumptions that we will make are: (a) The illumination used is coherent (b) signal-to-noise ratio (SNR) is infinite and (c) the length of the hologram or spatial light modulator is infinite.

The problem breaks down into two parts: representing the input and output collections in a useful form and determining the (KL) distance between them. The familiar plane-wave representation is adopted for both the input and output. Assume that the input is a collection of N plane-waves each propagating in a specific direction. The input collection may be represented as

$$\sum_{i=1}^{N} a_i e^{j2\pi f_i x} \qquad 3\text{-}9$$

The grating has a spatial frequency f_g and the wavelength of the light being used is λ. The set of plane waves produced as a consequence of the interaction between the input set of plane-waves and the grating may be represented as

$$\sum_{i=1}^{N} b_i e^{j2\pi (f_i + m\lambda f_g) x} = \sum_{i=1}^{N} (b_i e^{j2\pi m\lambda f_g x}) e^{j2\pi f_i x} \qquad 3\text{-}10$$

Eq 3-16 follows from the familiar relationship between the spatial frequencies

of the input and the output which is given below.

$$f_{out} = mf_g + f_{in} \tag{3-11}$$

where $m = \ldots -3, -2, -1, 0, 1, 2, 3, \ldots$ is the order of diffraction.

In order to determine the KL distance between the input and output plane wave distributions, mathematical objects known in quantum mechanics as *density matrices* [31, 32, 33] have to be introduced. The elements of a density matrix are all possible bilinear products of the coefficients of an expansion in some basis. As an example, consider Eq 3-15, where the input to the grating is represented as an expansion in a plane-wave basis. The density matrix corresponding to the input would then be determined by the outer product given below.

$$\rho = \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ a_N \end{bmatrix} \begin{bmatrix} a_1^* & a_2^* & . & . & a_N^* \end{bmatrix} \tag{3-12}$$

where * indicates complex conjugation. The matrix written out in full is

$$\rho = \begin{bmatrix} a_1 a_1^* & a_1 a_2^* & & a_1 a_N^* \\ a_2 a_1^* & a_2 a_2^* & & a_2 a_N^* \\ & & & \\ a_N a_1^* & a_N a_2^* & & a_N a_N^* \end{bmatrix} \tag{3-13}$$

The diagonal elements of the density matrix represent the energies corresponding to the different plane wave states. In quantum mechanics, the diagonal elements would represent the probabilities of photons being found in the particular states. Matrices with similar properties are also seen in digital signal processing, especially in spectral analysis. The use of density matrices

permits compact expressions and makes it easier to visualize more complex situations.

The relative entropy of two states ρ and σ is defined [34] as

$$D(\rho \| \sigma) = trace(\rho(\log_2\rho - \log_2\sigma)) \text{ bits} \qquad 3\text{-}14$$

where *trace* is the sum of the diagonal elements of the resulting matrix and is usually abbreviated to *tr*. It can be shown that this definition is equivalent to Eq. 3-2.

Both the ingredients that are required to compute the information content of a fringe pattern are available. They are now used to determine the information required to accomplish the change of state of a single plane wave of amplitude a_1 and spatial frequency f_1 impinging on a grating of spatial frequency f_g to produce output plane waves of amplitude b_m and spatial frequency $(f_1 + m\lambda f_g)$. The density matrices of the input and output states are simply scalars $|a_1^2|$ and $|b_m|^2$ respectively. The KL distance is then given by

$$D(input \| output) = \sum_{all\ m} |a_1|^2 \log_2\left(\frac{|a_1|^2}{|b_m|^2}\right) \text{ bits} \qquad 3\text{-}15$$

Clearly, this is a function of the energies in the two states and depends on the logarithm of the ratio of the energies. The summation is carried out over all m because a single input wave produces several output waves. This situation may be interpreted as a single wave impinging on several different gratings. At first glance, it appears that the amount of information transferred is independent of the grating spatial frequency. The dependency of the KL distance on the grating spatial frequency is quite subtle. In Eq. 3-15, the summation is carried out over all m. How many values of m are there? There are as many values of m as there are *non-evanescent* output waves. Rewriting Eq. 3-11 in terms of the angles of propagation, we have

$$\sin(\theta_{out}) = m\lambda f_g + \sin(\theta_{in}) \qquad 3\text{-}16$$

Non-evanescent output waves require that $\theta_{out} \leq \pi/2$. It follows that

$$m \leq \left(\frac{1 - \sin(\theta_{in})}{\lambda f_g} \right) \qquad 3\text{-}17$$

Because the number of output waves has to be an integer, the maximum value of m is given by

$$m_{max} = \left\lfloor \left(\frac{1 - \sin(\theta_{in})}{\lambda f_g} \right) \right\rfloor \qquad 3\text{-}18$$

where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. Therefore, the number of terms in the summation of Eq. 3-15 is quite strongly dependent on both the wavelength of light and the grating spatial frequency. Of course, Eq. 3-15 is not valid for purely sinusoidal gratings which produce only the zero and positive and negative first orders. To recapitulate, the ingredients required to compute the amount of information in a holographic fringe pattern are the input and output density matrices and Eq. 3-15.

A few qualitative observations about Eq. 3-15 are in order here. The KL distance depends on the reciprocal of the energies in each diffracted component. This has two implications. First, the smaller the output energy (for unit input energy), the greater the number of bits required. This conclusion seems to make sense when we take into account the fact that smaller output energies imply more precise control over the input wave. It takes more effort, in terms of bits, to achieve precise control. The second point to note has to do with the number of output waves. A low spatial frequency grating produces several output orders and the energies of those orders are smaller than when a high spatial frequency grating is used. Thus, the KL distance is larger for a lower spatial frequency grating than for a high spatial frequency grating — a rather surprising conclusion. The explanation of this result lies in realizing that low spatial

frequency gratings produce several non-evanescent orders. In order to *resolve* — i.e., physically distinguish — all these orders, a greater amount of information is required. The classical connection between information and resolution [35, 36, 37] emerges naturally from this argument. Another connection to classical information theory is also manifest in Shannon's analysis of entropy loss in linear filters [28].

In order to test the above theory on real gratings, a few examples [38] are presented below.

A piece of clear glass: The transmittance function of a piece of clear glass is unity. It simply lets all the light pass through. In this case, the output plane wave has exactly the same amplitude and direction as the input plane wave. Therefore the information transferred to the input plane wave by the piece of glass must be zero. Substituting the input and output amplitudes into Eq. 3-15 and noting that there are no diffracted orders, we have:

$$D = 1\log_2\left(\frac{1}{1}\right) = 0 \text{ bits} \qquad 3\text{-}19$$

A piece of neutral density glass: This piece of glass attenuates the input light but leaves the direction intact. If we assume that the amplitude transmittance is t_0, then we obtain:

$$D = 1\log_2\left(\frac{1}{t_0^2}\right) \text{ bits} \qquad 3\text{-}20$$

Sinusoidal amplitude grating: The transmittance function of a sinusoidal amplitude grating may be written as

$$t(x) = t_0 + \Delta t \sin(2\pi f_x) \qquad 3\text{-}21$$

where t_0 is the bias term and Δt is the maximum excursion from the mean value. It is a well known fact that the energies of the diffracted components are proportional to the coefficients of the Fourier expansion of the transmittance function for thin gratings. Using this fact, we obtain the coefficients of the zero

and ±1 orders. They are given by t_0, $\Delta t/2$, and $\Delta t/2$ respectively. Substituting these values into Eq. 3.15 and assuming that the input plane wave has unit amplitude, we have

$$D(input \| output) = \log_2\left(\frac{1}{t_0^2}\right) + 2\log_2\left(\frac{1}{(\Delta t/2)^2}\right) \qquad 3\text{-}22$$

For the case where $t_0 = 0.5$ and $\Delta t = 0.5$, we compute the KL distance to be 10 bits.

Square-wave amplitude grating: The transmittance function of a square-wave transmission grating is represented by

$$t(x) = t_0 + \Delta t \,\mathrm{sq}(2\pi f_x) \qquad 3\text{-}23$$

where sq(.) represents a square wave with a duty cycle of 50%. The coefficients of the Fourier expansion of this transmittance function are

$$t_0, \frac{2\Delta t}{\pi}, 0, \frac{2\Delta t}{3\pi}, 0, \frac{2\Delta t}{5\pi}, 0, \ldots\ldots\ldots\ldots \qquad 3\text{-}24$$

Substituting these values into Eq. 3-15, we have

$$D = \log_2\left(\frac{1}{t_0^2}\right) + \sum_{odd\ i}\left(\log_2\left(\frac{1}{(2\Delta t/\pi)^2}\right) + \log_2(i^2)\right) \text{ bits} \qquad 3\text{-}25$$

Eq. 3-25 is rewritten below to make some interesting facts apparent.

$$D = \log_2\left(\frac{1}{t_0^2}\right) + \sum_{odd\ i}\log_2\left(\frac{1}{(2\Delta t/\pi)^2}\right) + \sum_{odd\ i}\log_2(i^2) \text{ bits} \qquad 3\text{-}26$$

The first term in Eq. 3-26 pertains to *zero-order* light. The second may be

interpreted as relating to the *amplitudes* of all diffracted orders and the final term may be interpreted to be the *direction-related* component of the KL distance. It is clear from the definition of the direction related component that it takes more information to deviate the incoming plane wave through a larger angle — an observation that is consistent with the quantum mechanical interaction picture in acousto-optic modulators. By expressing the KL distance as the sum of three components, we can decouple amplitude and directional information. Each of these quantities may be treated independently — a feature that will come in handy when more complex gratings are analyzed.

Square-wave phase grating: The transmittance function of a square-wave phase grating is as follows:

$$\phi(x) = \phi_0 + \Delta\phi \,\mathrm{sq}(2\pi f_x) \qquad 3\text{-}27$$

where ϕ_0 is the constant bias and $\Delta\phi$ is the maximum excursion from ϕ_0. The coefficients of the Fourier expansion of this transmittance function are given by:

$$\cos(\Delta\phi), \frac{2}{\pi}\sin(\Delta\pi), 0, \frac{2}{3\pi}\sin(\Delta\pi), 0, \frac{2}{5\pi}\sin(\Delta\pi), \ldots\ldots \qquad 3\text{-}28$$

Substituting these values into the now familiar Eq. 3-15, we get

$$D = \log_2\left(\frac{1}{\cos^2(\Delta\phi)}\right) + \sum_{odd\ i}\log_2\left(\frac{\pi^2}{4\sin^2(\Delta\pi)}\right) + \sum_{odd\ i}\log_2(i^2) \qquad 3\text{-}29$$

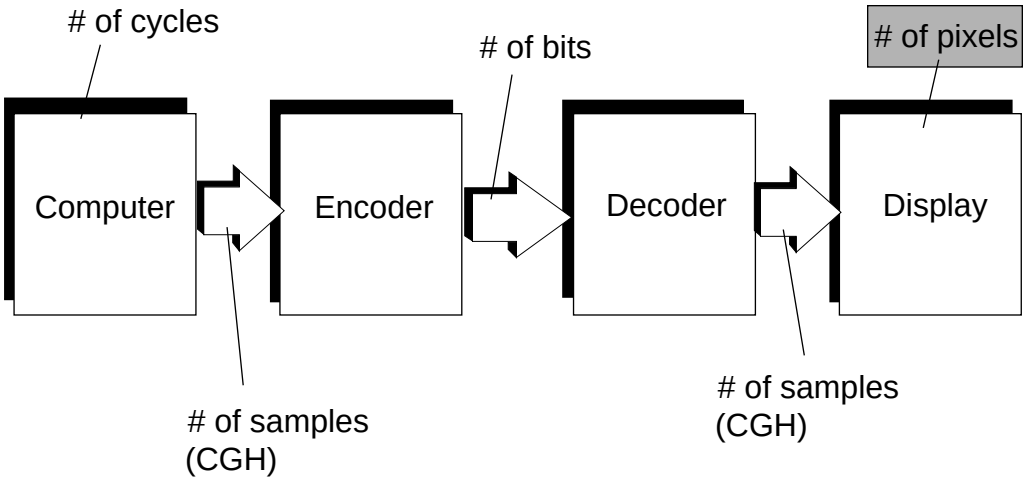
To summarize the salient points of this section: the two ingredients required to compute the amount of information that a grating imparts to an incoming plane wave are the input and output density matrices. Once these are determined, the Kullback-Liebler distance may be computed using Eq. 3-15. The result of the computation is the least number of bits required to encode grating transmittance function. The KL distance may be expressed as the sum of three independent components, each of which accounts for the zero-order term, amplitude dependence and direction dependence of the KL distance on the diffracted

components.

3.5 Summary

This chapter described a bottom-up approach to determining the information content of a hologram. In contrast to previous approaches, which treated the holographic fringe pattern as an image, the mechanism of information transfer in a fringe pattern was modeled using information theory principles. The predictions of the model appear to be consistent with the observations of previous researchers. The model is valid only under the three assumptions stated in section 3.4. Further work should include extending the model to incoherent sources of illumination and including the effects of noise and finite modulator length. The author is firmly convinced that a complete theory of information transfer needs to be formulated in order to arrive at practical methods of hologram computation for scalable displays.

4 Pixel aspect ratios in holographic displays



4 Pixel aspect ratios in holographic displays

4.1 The display

The display is the penultimate stop in the communication chain before holographic image information is presented to the human visual system. The decoded CGH is presented to the beam of light - which may be coherent or partially incoherent - via a spatial light modulator (SLM). Therefore, the display performs two important tasks - modulation and image presentation. In this chapter, we will be concerned exclusively with the physical format of presentation of information to the beam of light - that is, the complete description of the pixels of the SLM. A complete description entails specifying the size of each pixel and the dynamic range that each pixel should have in order to make the displayed image appear perceptually lossless.

The next section is an introduction to the generalized image relay system (GIRS), which provides the framework in which pixel sizes will be determined. Then, canonical geometries are presented for the spherical GIRS case and pixel sizes are determined. Cylindrical-optical extensions of the GIRS are presented in section 4.4. The final section discusses the utility of determining the pixel sizes and dynamic range requirements.

4.2 The generalized image relay system

The GIRS is a simple two-lens relay system [39] as shown in Figure 2.1. L1 and L2 are identical spherical lenses which are separated by twice their focal length F. L1 is the *objective lens* and L2 is the *field lens*. L1 forms a real image of the three-dimensional scene in space. The primary function of L2 is to prevent vignetting of the image. It has no effect on the power of the system *per se*, but it bends rays which would otherwise leave the system back towards the axis. Another important function of L2 is that it corrects simple geometric distortions so that longitudinal and lateral magnifications of the image are independent of z-location. This statement may be proved simply. The effective focal length $f_{effective}$ of a pair of thin lenses of focal lengths f_1 and f_2 separated by a distance d is given by

$$\frac{1}{f_{effective}} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2} \qquad 4\text{-}7$$

For our case, $f_1 = F$, $f_2 = F$, and $d = 2F$. This gives:

$$f_{effective} = \infty \qquad 4\text{-}8$$

which means that the magnification of the system is unity and the image is the same size as the object and inverted regardless of the location of the object along the axis. The connection between the GIRS and holography is the subject of the rest of this section.

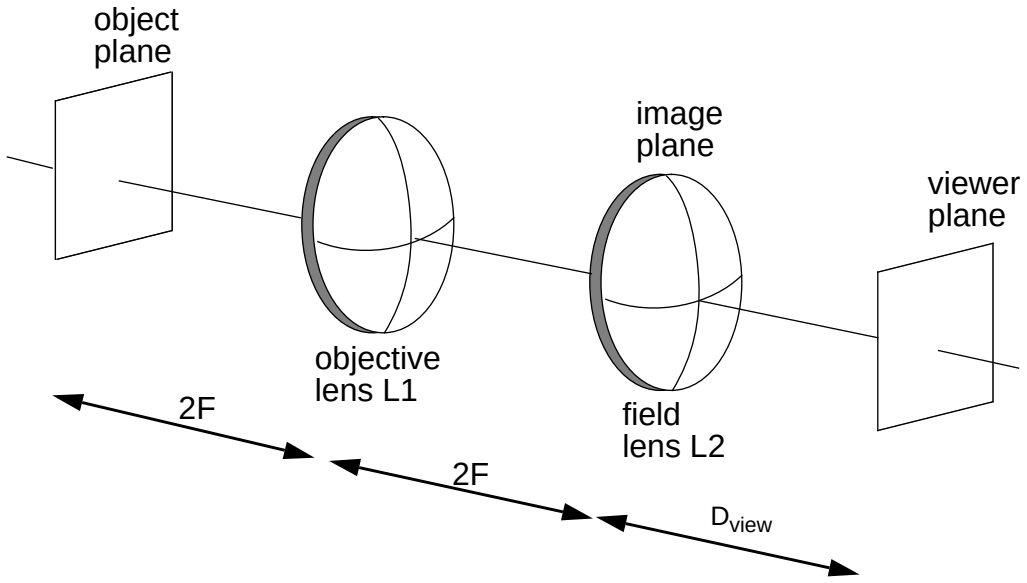


Figure 4.1: The generalized image relay system

To model a hand-held holographic image viewing configuration, the images are taken to be 100 mm on a side and the viewing distance is 500 mm. This implies that the lenses have a focal length of 250 mm, and a diameter of 141 mm. The images in this configuration are approximately the same size as the images in the Mark II system. The f-number of the lenses is approximately 1.8. The magnification is taken to be unity without loss of generality. Consider inserting a holographic plate at some plane in the system and recording a hologram in that plane. All components downstream of the plate may be removed in the recording process. The image may then be reconstructed from that hologram by "playing it back" through the rest of the system. This is shown in Figures 4.2(a) and 4.2(b). If the hologram were recorded just before or after the field lens L2, it

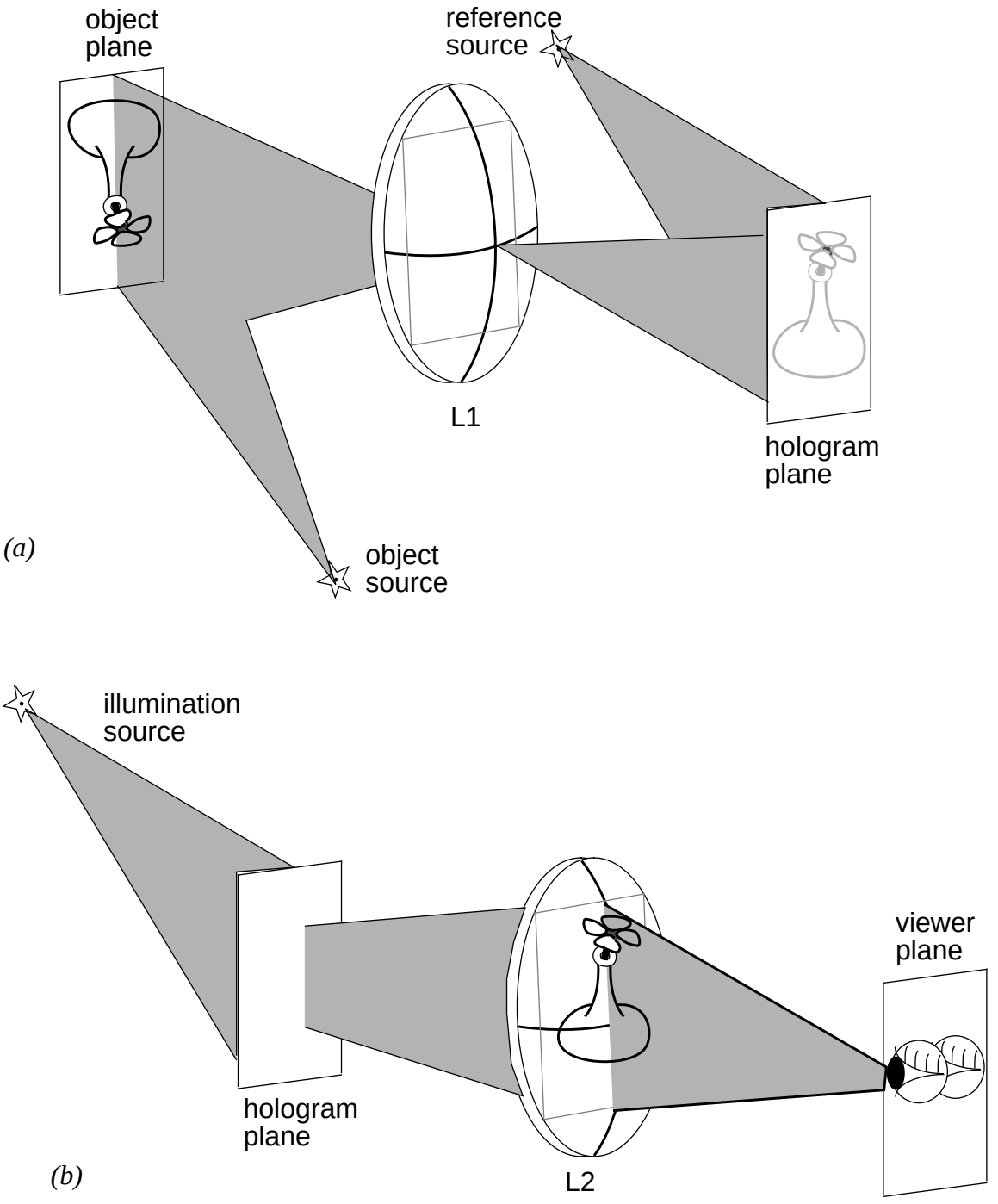


Figure 4.2: (a) Recording and (b) reconstruction of a hologram in a GIRS

would bisect the three-dimensional image of the scene formed by L1. This is the equivalent of an *image-plane hologram* in which the viewer looks *at* the hologram. The other extreme, a *viewer-plane hologram*, can be recorded by putting the hologram plane just before or after the objective lens L1. Lens L2 would image the hologram to the viewer's plane, who would look *through* it at the image in the plane of L2. The location of the hologram plane between L1 and L2 is a free variable — a knob — using which the hologram may be recorded as an image-plane hologram, a viewer-plane hologram, or anything in between. Another degree of freedom is the location of the reference source.

The GIRS may also be used to model and analyze *horizontal parallax only* (HPO) imaging geometries. This is shown in Figure 4.3. Reduction of vertical parallax is achieved by using a horizontal slit in the plane of L1, much like in a rainbow hologram. The slit permits only a side-to-side range of perspectives to enter the system. The view-zone is now an aerial image of the slit. A vertical diffusing element must be provided after L2 in order to make the view-zone usable. The introduction of the slit brings with it *astigmatism* - particularly in deep images. Astigmatism is a feature of *any* HPO imaging scheme, holographic or otherwise. It occurs because vertical information is always

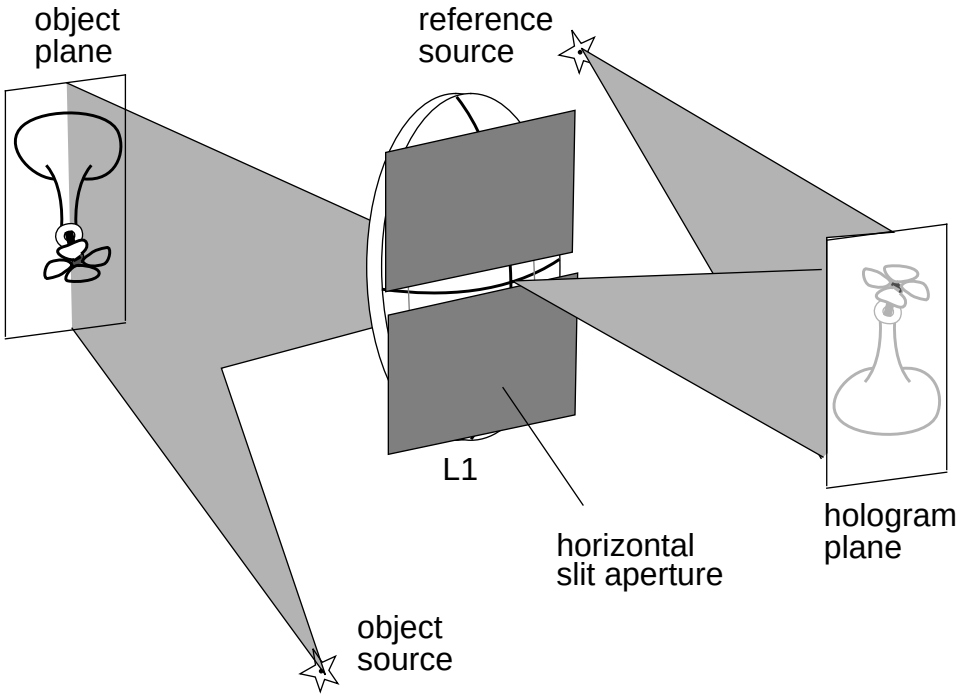


Figure 4.3: HPO imaging in a GIRS

centered at the surface of the diffuser while horizontal information is centered at the apparent viewing distance. Astigmatism limits the usable depth of the image.

As was mentioned before, one degree of freedom is the location of the reference source. The recording of an image-plane HPO hologram is shown in Figure 4.4. The hologram plane is located just before L2 and the reference source is placed at the end of the slit. In optical holography, the source would be placed at least the width of the slit from its end, but this is not a restriction in computer-

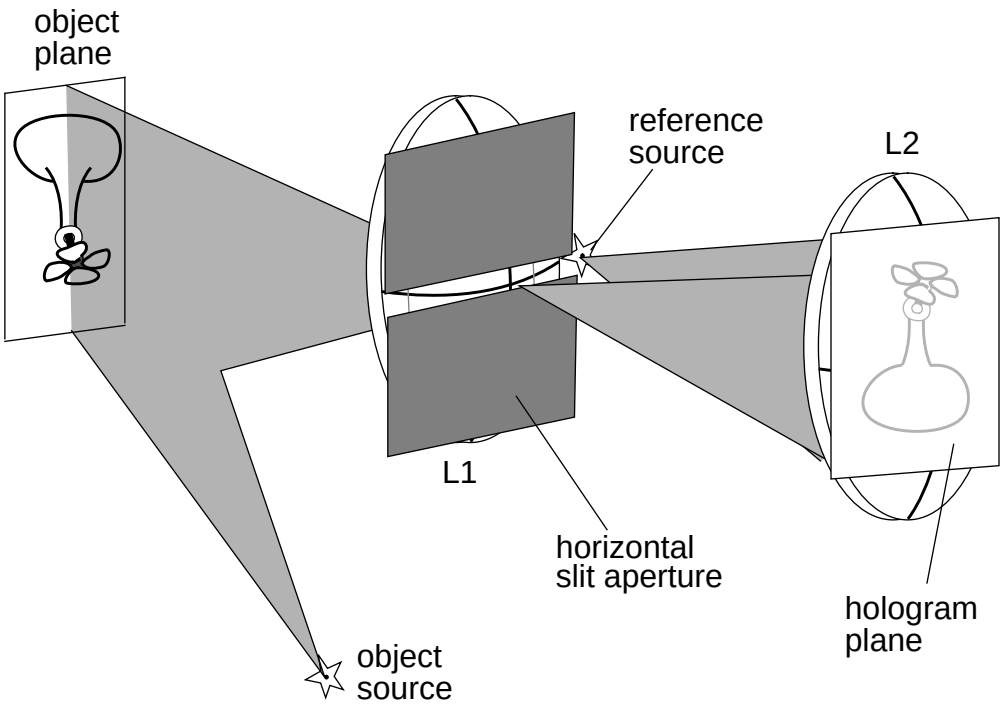


Figure 4.4: Reference source location. In computational holography the source can be located at the edge of the slit.

generated holography, because the “halo” term can explicitly be excluded in the computation. The position of the reference source is chosen so as to keep the maximum spatial frequency of the hologram transmittance pattern approximately constant over the hologram area. If the reference source were in a different location or produced plane waves, the maximum spatial frequency would vary across the hologram area. The consequences of varying spatial frequency was the subject of Chapter 2. In this chapter, uniform sampling is assumed. In the following sections, we take a closer look at canonical or prototype uniformly-sampled computed hologram geometries and determine pixel aspect ratios for each case.

4.3 Canonical geometries: spherical case

Canonical geometries refer to either image-plane or viewer-plane hologram configurations [40]. As was noted before, there are two degrees of freedom involved: the location of the hologram plane and geometry of the reference beam. In this section, we explore all possible combinations of hologram location and reference beam geometry when spherical lenses are used in the GIRS. The cylindrical lens case is examined in the following section. In the interest of brevity, only the image-plane case will be considered completely; the viewer-plane case will be outlined and only results will be stated.

Full-parallax image-plane holograms: Table 4-1 shows the parameters for all possible canonical configurations. The reference source locations and hologram locations are indicated in Figure 4.5. For example, one canonical geometry for a

Table 4-1: Computed hologram parameters

Hologram type	Hologram location	Reference source location	Slit
Full-parallax image-plane	L2	1,2	✘
HPO image-plane	L2	1,3	✔
Full-parallax viewer-plane	L1	4,5	✘
HPO viewer-plane	L1	4,5	✔

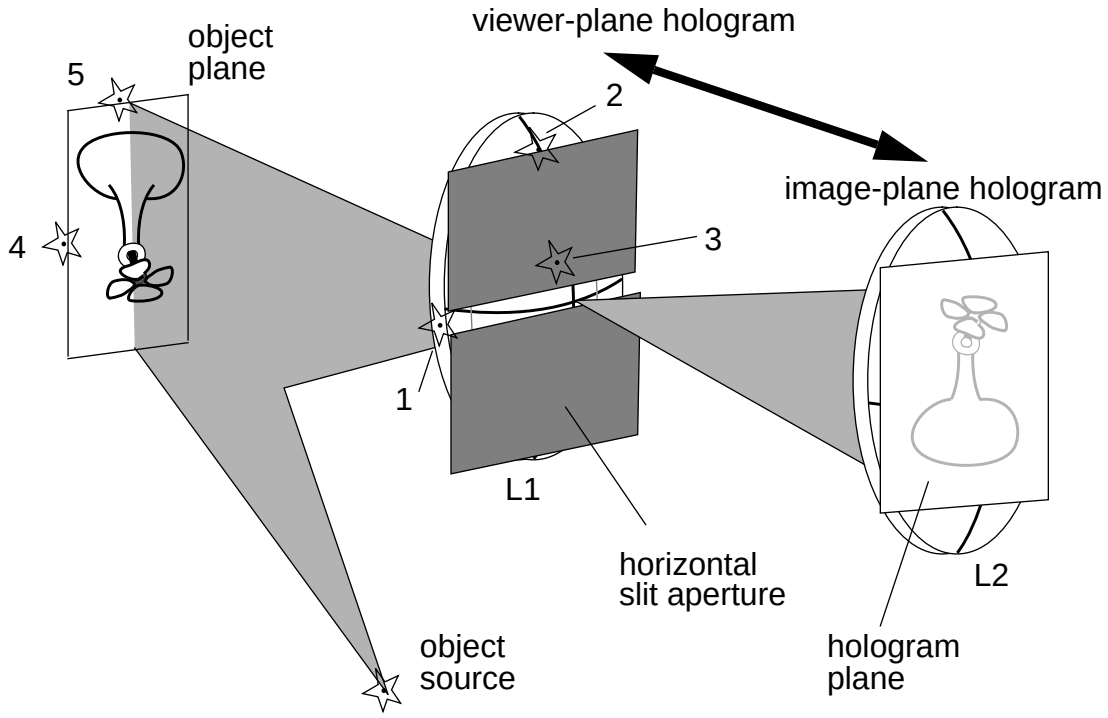


Figure 4.5: Hologram and reference source locations

full-parallax image-plane computed hologram would consist of the following: a hologram plane located at L2, a reference source located at location 1 in Figure 4.5, and no slit. This is indicated by the ✘ in the 'slit' column. A second canonical configuration would require the reference source to be placed at location 2 in Figure 4.5.

The following analysis is valid for the full-parallax image-plane case with the reference source at location 1. The analysis for all other cases is similar and will not be presented. We begin by noting that the pixel dimensions depend only on the maximum sampling rates which in turn depend only on the maximum spatial frequencies that are recorded on the hologram. in the x and the y directions. Let us label these maximum spatial frequencies $f_{x,\,max}$ and $f_{y,\,max}$ respectively. In order to simplify things, consider the variation of f_x vs. f_y as two point sources interfere with each other in the hologram plane. One of the points is at location 1 while the other point traverses the semi-perimeter of a square aperture as shown in Figure 4.6. Clearly, $f_y = f_{y,\,max}$ when the second point is directly

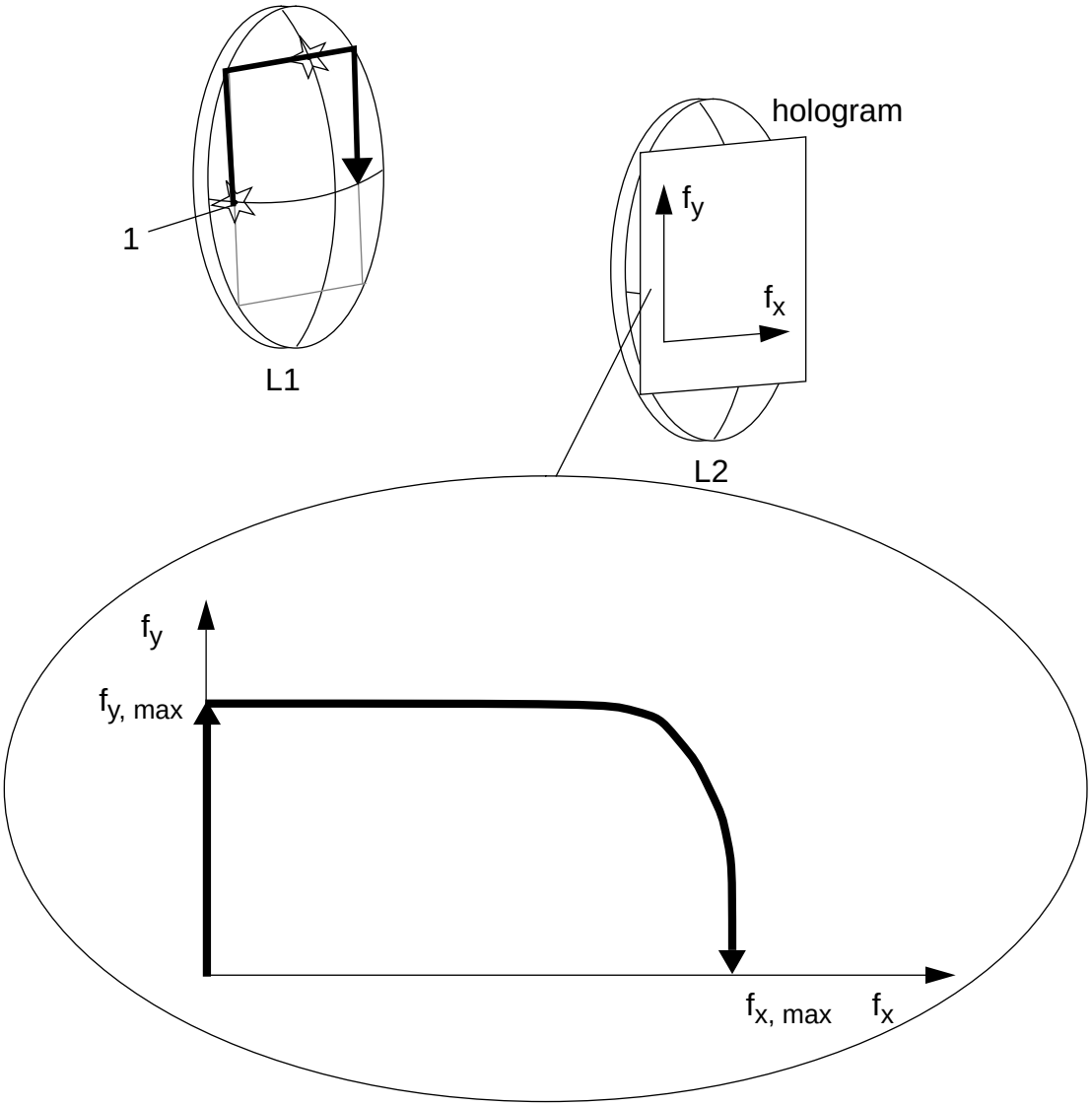


Figure 4.6: Two points on the edge of the aperture - one stationary and the other moving. The lower figure shows the variation in the spatial frequencies.

above the first point i.e., directly above location 1. This observation simplifies the computation of the maximum spatial frequencies considerably. Let $\theta_{H,\,max}$ and $\theta_{V,\,max}$ be the maximum angles subtended at the hologram plane in the *x* and *y* directions respectively and *D* be the length of the square aperture. Then the following equations are true.

$$\theta_{H,\,max} = 2\tan^{-1}\left(\frac{D}{4F}\right)$$

$$\theta_{V,\,max} = \tan^{-1}\left(\frac{D}{4F}\right) \qquad \text{4-9}$$

where F is the focal length of either L1 or L2. Therefore, the maximum spatial frequencies in both directions are given by:

$$f_{x,\,max} = \frac{\sin(\theta_{H,\,max})}{\lambda}$$
$$f_{y,\,max} = \frac{\sin(\theta_{V,\,max})}{\lambda} \quad \text{line pairs per mm} \qquad 4\text{-}10$$

where λ is the wavelength of light used to record the hologram. Recognizing that the hologram transmittance pattern must be sampled at twice the maximum spatial frequency in either direction, following the prescription of the Shannon-Nyquist sampling theorem, we determine the number of samples in each millimeter of the hologram. The dimensions of each pixel follow in a straightforward way from the number of samples per unit length. Let us label the pixel dimensions p_x and p_y. Then:

$$p_x = \frac{\lambda}{2\sin(\theta_{H,\,max})}$$
$$p_y = \frac{\lambda}{2\sin(\theta_{V,\,max})} \quad \text{mm} \qquad 4\text{-}11$$

If the dimensions of the hologram are $P \times Q$ mm, then the total number of pixels required is

$$N_{full-parallax} = \left(\frac{4}{\lambda^2}\right)(PQ)\sin(\theta_{H,\,max})\sin(\theta_{V,\,max}) \qquad 4\text{-}12$$

This is the *space-bandwidth product* (SBWP) or the *Smith-Lagrange optical invariant* of the hologram. Clearly, the SBWP depends on the hologram area and the angles subtended at the hologram, as expected. *Any* full parallax hologram recorded in the GIRS will have exactly the same number of pixels. Evaluating (3) and (4) for a test case using F = 250 mm, D = 100 mm, and λ = 633 nm, we obtain p_x = 1.6μm, p_y = 3.2μm, and the total number of samples $N_{full-parallax} = 1.96 \times 10^9$. This is a prohibitively large number of samples and was one of the primary factors in discouraging holographic video research in the early days of holography. The aspect ratio of these pixels is 1:2.

HPO image-plane holograms: In the previous case, we computed the pixel dimensions for each axis independently. This is a useful approach, because it simplifies the computation for the HPO case. The horizontal dimension of the pixel remains exactly the same. The only change is in the vertical dimension, because we don't require diffraction in the vertical direction. If we require L lines in the hologram, then the vertical dimension of each pixel becomes $p_y = Q/L$ mm. For our example, p_y becomes 390μm, giving an aspect ratio of 1:244. In this case, we require pixels which are much taller than they are wide, which matches our intuition about HPO holograms. Holographic resolution is required in the horizontal direction but not in the vertical direction.

Table 4-2: Full-parallax holograms

Hologram type	Reference source location	Pixel dimension (μm)	Pixel aspect ratio
Image-plane	1	1.6×3.2	1:2
	2	3.2×1.6	2:1
Viewer-plane	4	1.6×3.2	1:2
	5	3.2×1.6	2:1

Table 4-3: HPO holograms

Hologram type	Reference source location	Pixel dimension (μm)	Pixel aspect ratio
Image-plane	1	1.6×390	1:244
	3	3.2×390	1:122
Viewer-plane	4	1.6×3.2	1:2
	5	3.2×1.6	2:1

If the reference source is now moved to position 2 (Figure 4.5), the pixel aspect ratios are simply transposed for the full parallax case. For the HPO case, they become $3.2\mu m \times 390\mu m$, giving an aspect ratio of 1:122. A similar analysis may be performed for the viewer-plane case. Pixel aspect ratios for all canonical MPH geometries are presented in Tables 4-2 and 4-3 below. At first glance, it appears that the SBWP is not the same for the image-plane and viewer plane cases in Table 4-3, because the pixels are much smaller in the viewer-plane case. This apparent violation of SBWP constancy may be explained by the fact that the viewer-plane HPO hologram has *a smaller area* than a corresponding image-plane HPO hologram The fact that the SBWP is invariant in both cases can be used to determine the slit height. In this case, the slit height is 0.82 mm, which exactly matches the value derived from Fourier considerations [39].

4.4 Canonical geometries: cylindrical case

It is clear that the variety of pixel aspect ratios changes considerably when HPO imaging is taken into account. In this section we examine cylindrical optical extensions of the GIRS and determine the pixel aspect ratios. Our motivation for studying the cylindrical case is the following. First, the variety in pixel aspect ratios is broadened. Second, there is an additional degree of freedom when cylindrical lenses are used, which provides the ability to control the amount of astigmatism. Finally, the dynamic range requirement on the samples may be relieved in this case, although this problem is deferred to future studies. We will consider the most useful case in which the image of the modulator is in the image-plane.

Consider Figures 4.7(a) and 4.7(b). Figure 4.7(a) depicts the vertical subsystem while Figure 4.7(b) depicts the horizontal subsystem. Our goal is to set up a geometry such that the image is anastigmatic and the image of the hologram is astigmatic. This means that horizontal and vertical image information are focused at lens LV4. The hologram plane is located at lens LV2. Horizontal hologram information is focused at LV4 while vertical hologram information is focused at the viewer-plane - the hologram is imaged to *two* different planes. From the discussion in section 4.2, it is clear that this arrangement produces an image-plane hologram as far as horizontal hologram information is concerned and a viewer-plane hologram for the vertical hologram information. This is clear from Figures 4.7(a) and 4.7(b) where the viewer looks *at* the horizontal hologram and *through* a vertical hologram at an anastigmatic image. There are

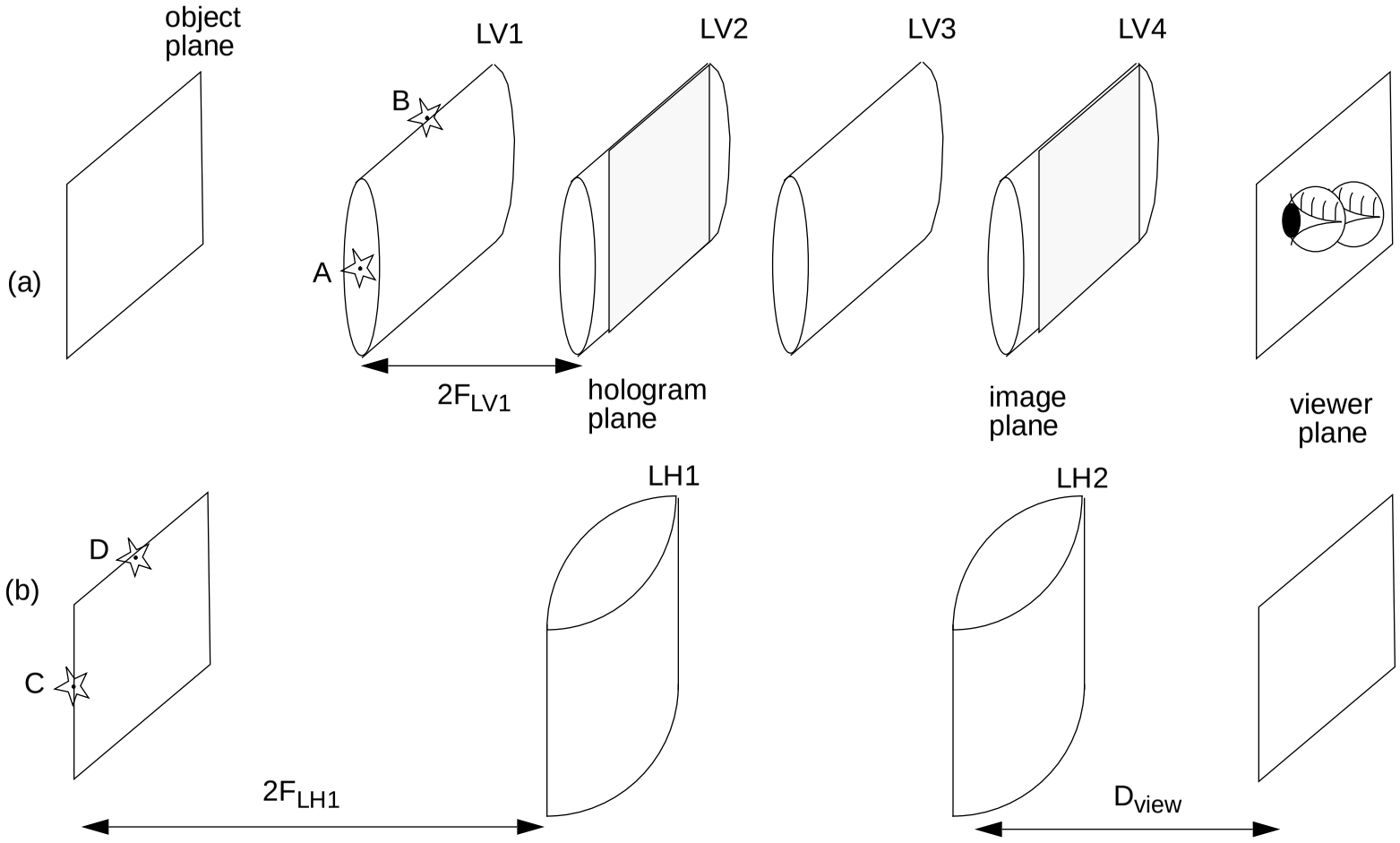


Figure 4-7: (a) The vertical subsystem of the cylindrical GIRS and (b) the horizontal subsystem. Note that the image is anastigmatic while the holograms are spatially separate. A, B, C, and D are the locations of the reference sources.

four possible reference source locations - A, B, C, D - as shown in Figures 4.7(a) and 4.7(b). We use an analysis similar to the one in section 4.3 to determine pixel aspect ratios for both the full parallax and HPO cases.

The results are summarized in Tables 4-4 and 4-5. It must be noted that lens LH2 should have a different focal length than lens LH1. If we assume that the viewing distance is D_{view}, then we have the following relationship for the required focal length.

$$F_{LH2} = \frac{2D_{view}F_{LH1}}{2F_{LH1} + Dview} \qquad 4\text{-}13$$

Continuing our previous example, with D_{view} = 500 mm, we have F_{LH2} = 250 mm. All values in the tables below pertain to computed holograms which mimic hand-held holograms as mentioned in section 4.2.

Table 4-4: Full-parallax holograms - cylindrical case

Reference source location	Pixel dimension (μm)	Pixel aspect ratio
A	1.6×1.8	1:1.125
B	0.852×1.6	1:1.880
C	1.6×0.823	1:0.514
D	1.6×0.852	1:0.532

For the HPO case, we interpose a horizontal slit aperture at lens LV2 that allows only a limited vertical perspective. A vertical diffusing element is located in the plane of lens LV4. In this case, the vertical pixel dimensions are determined by the number of lines required in the hologram. For our example, the number of lines is 256, as before. We also use the same slit width as before: 0.82 mm. This leads to the entries in Table 4-5. For purposes of comparison, we note that some of the newest LCDs from Epson have aspect ratios between 1:1 and 1:3 with common pixel dimensions being on the order of tens of microns.

Table 4-5: HPO holograms - cylindrical case

Reference source location	Pixel dimension (μm)	Pixel aspect ratio
A	1.6×3.2	1:2
B	0.852×3.2	1:3.755
C	1.6×3.2	1:2
D	1.6×3.2	1:2

4.5 Discussion

In this chapter, we determined the pixel sizes required in various canonical near-field geometries. One fact that clearly emerged from this exploration was that cylindrical-optical GIRS configurations allow for a greater variety in pixel sizes than spherical-optical ones. This variety stems from the fact that there are two subsystems working together, and their relative position introduces a free variable which is reflected in the pixel dimension choices.

Another novel concept was the introduction of anastigmatic images and astigmatic horizontal and vertical holograms. A related approach was followed by Fritzler and Marom [25] in their quest to reduce the information content of a holographic image. It must be mentioned that Vanderlugt [12] also performed a careful analysis of the spatial frequencies at different planes in the system. He concluded that by using pixels with variable aspect ratios over the hologram plane it was possible to represent near-field holograms with the same number of samples as required for far-field holograms. This chapter dealt with samples which have the same dimensions all over the hologram plane, keeping a practical system in mind.

There are, however, various other consequences and applications of this analysis. First, if a modulator with a given pixel aspect ratio is available, this analysis useful in deciding what optics must be used so as to make full use of the SBWP of the modulator. The difference between pixel aspect ratio given and that required can be used to determine the power of the optical system used to image the modulator to a different plane. Second, the analysis may be carried out for any geometry with ease. This allows modulator designers to determine

the required aspect ratios in advance while designing custom SLMs for holographic imaging.

Finally, having decided how big each pixel must be in a particular geometry, we must decide how much dynamic range or bit-depth each pixel must have. This is a question that requires careful consideration. A thorough survey of the literature in computer-generated holography revealed that no careful analysis of the bit-depth requirements has been performed for near-field display holograms. This is the subject of studies in the near future.

5 Summary, conclusions, and future directions

5 Summary, conclusions, and future directions

5.1 Summary and conclusions

In this section, the achievements of the foregoing three chapters are highlighted. The goal of this thesis was to characterize the various elements of the information landscape in a holographic display with the hope that such a characterization would provide insight into eliminating redundancy in computed fringe patterns. The greater goal was to build a firm foundation on which to base the scale-up process of current holographic display systems. The contribution of each chapter is briefly summarized below.

Nonuniform sampling of fringe patterns: An algorithm to reduce the total number of samples in a holographic fringe pattern by using nonuniform sampling was presented. A mathematical proof of the algorithm was developed and justified on both visual and perceptual grounds. To the best of the author's knowledge, this is the first ever time such an algorithm has been used to compute display holograms. The issues/trade-offs involved were presented and the performance of the algorithm was demonstrated on the Mark II MIT Holographic Video System. The results clearly indicate the lossless nature of the algorithm.

Information transfer in spatial light modulators: In chapter 3 an attempt was made to determine the minimum number of bits that would be required to encode a fringe pattern. The approach was firmly grounded in information theory and the expressions derived were tested on real gratings to determine their validity. The numbers obtained seemed reasonable and various limiting cases gave satisfactory answers. The method is amenable to experiment in that predictions may be made and tests may be carried out.

Pixel aspect ratios in holographic displays: General methods were developed in chapter 4 to determine the pixel aspect ratios required in SLMs for use in holographic displays. The advantage of knowing the required aspect ratio for a particular geometry is that any SLM may be optically transformed to match the requirements — thus fully utilizing the available bandwidth. Both spherical and the more general cylindrical case were considered. A method that produced

anastigmatic images with astigmatic horizontal and vertical holograms was presented.

5.2 Future work

Several new avenues of research emerge from the work presented in the previous three chapters. These directions are summarized below.

Nonuniform sampling of fringe patterns: This is one area where a great deal of work is possible. First, faster ways have to be found to implement the algorithm. This will enable interactive displays to be built. Currently, the performance of the algorithm in terms of speed is not satisfactory, because of the large amount of interpolative filtering that has to be performed. The ideal solution to this problem would be an optical one, in which case the performance would improve dramatically. Second, it was suggested in section 2.6 that nonuniform segment widths could be used to advantage. Once again, such a scheme would require a fast decompression method to be available. The greatest gains, however, will be possible when a bandpass sampling algorithm for CGHs is implemented. This algorithm is the subject of research at the current time.

Information transfer in spatial light modulators: This is another area where a significant progress can be made. The contribution of chapter 3 is only the tip of the iceberg. A comprehensive analysis of the various representations of information and their relationship to one another will be the cornerstone of the scale-up process of holographic displays. At the current time, the question "*what is the information content of a hologram?*" remains almost wholly unanswered.

It is important to break away from the attitude of treating a holographic fringe pattern as an image. It must be pointed out that the first estimates of the number of samples in a hologram relied on uniform Nyquist sampling and treating the holographic fringe pattern as an image. It is the author's firm conviction that there is more structure to a holographic fringe pattern than the foregoing treatment would have one believe. Several researchers are currently working on the problem of determining the degrees of freedom of optical systems and channels. Similar analyses must be performed on holographic displays in order to gain insight into the modalities of information propagation in such displays.

Just as the Fast Fourier Transform algorithm changed the face of digital signal processing by exploiting structure in the computation, an algorithm that exploits the structure of holographic fringe patterns will one day be the basis for real-time, life-size, interactive holographic displays.

References

References

[1] Gabor, D., *A New Microscopic Principle*, Nature, 161, 777-8, (1948).

[2] Kozma, A. and Kelly, D. L., *Spatial filtering for detection of signals submerged in noise*, Applied Optics 4, 387 (1965).

[3] Brown, B. R., and Lohmann, A. W., *Complex spatial filtering with binary masks*, Applied Optics 5, 967 (1966).

[4] Kollin, J. S., Benton, S. A, and Jepsen, M. L, *Real-Time Display of 3-D Computed Holograms by Scanning the Image of an Acousto-Optic Modulator,* in: G.M. Morris, ed., SPIE Proc. Vol. #1136, Holographic Optics II: Principles and Applications (1989).

[5] Underkoffler, J. S., *Development of Parallel Processing Algorithms for Real-Time Computed Holography,* SB Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, (1988).

[6] St. Hilaire, P., *Real time holographic display: Improvements using higher bandwidth electronics and a novel optical configuration,* SM Thesis, Media Arts and Sciences Section, Massachusetts Institute of Technology, (1990).

[7] St-Hilaire, P., Benton, S. A., Lucente, M., Sutter, J. D., and Plesniak, W. J., *Advances in Holographic Video*, in S. A. Benton, ed., SPIE Proceedings Vol. #1914: Practical Holography VII: Imaging and Materials (1993).

[8] St. Hilaire, P., *Scalable Optical Architectures for Electronic Holography,* Ph.D Thesis, MIT Program in Media Arts and Sciences, Massachusetts Institute of Technology, (1994).

[7] Bove, V. M. Jr., and Watlington, J., *Cheops: A data-flow processor for real-time video processing*, MIT Media Laboratory Technical Memo (1993).

[10] Lucente, M., *Diffraction-Specific Fringe Computation for Electro -Holography,* Ph.D Thesis, Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology (1994).

[11] Sutter, J., *Viewer-Plane Experiments with Computed Holography with the MIT Holographic Video System,* SM Thesis, MIT Program in Media Arts and Sciences, Massachusetts Institute of Technology, (1994).

[12] Vanderlugt, A., *Optimum Sampling of Fresnel Transforms,* Applied Optics, Vol. 29, No. 3, August 1990.

[13] Brown, B. R., and Lohmann, A. W., *Computer Generated Binary Holograms,* IBM Journal of Research and Development, 13, 160-167, (1969).

[14] Haskell, R. E., and Culver, B. C., *New Coding Technique for Computer-Generated Holograms,* Applied Optics, 11, 2712-14 (1972).

[15] Lee, W. H., *Sampled Fourier Transform Hologram Generated by Computer,* Applied Optics, 9, 639-43 (1970).

[16] Burckhardt, C. B., *A Simplification of Lee's Method of Generating Holograms by Computer,* Applied Optics, 9, 1949 (1970).

[17] Lesem, L. B., Hirsch, P. M., and Jordan Jr., J. A., *The Kinoform: A New Wavefront Reconstruction Device,* IBM Journal of Research and Development, 13, 150-5 (1969).

[18] Chu, D. C., Fienup, J. R., and Goodman, J. W., *Multi-emulsion, On-axis, Computer Generated Hologram,* Applied Optics, 12, 1386-8, (1973).

[19] Waters, J. P., *Three-dimensional Fourier Transform Method for Synthesizing Binary Holograms,* Journal of the Optical Society of America, 58, 1284-8, (1968).

[20] Fienup, J. R., *Iterative Method Applied to Image Reconstruction and to Computer-Generated Holograms,* Optical Engineering, Vol. 19, No. 3, 297-305, March 1980.

[21] Oppenheim, A. V., and Schafer, R. W., *Discrete-time Signal Processing*, Prentice Hall, (1989).

{22] Foley, J. D., et. al., *Computer Graphics: Principles and Practice*, Addison-Wesley, (1990).

[23] Haykin, S., *Communication Systems*, Wiley Eastern Limited, (1989).

[24] Marks II, R. J., *Introduction to Shannon Sampling and Interpolation Theory*, Springer-Verlag, (1991).

[25] Fritzler, D., and Marom, E., *Reduction of bandwidth required for high resolution hologram transmission*, Applied Optics, Vol. 8, No. 6, (1969).

[26] De Bitetto, D. J., *Bandwidth reduction of hologram transmission by elimination of vertical parallax*, Applied Physics Letters, Vol. 12, No. 5, 176-178, March 1968.

[27] Lin, L. H., *A method of hologram information reduction by spatial frequency sampling*, Applied Optics, Vol. 7, No. 3, March 1968.

[28] Shannon, C. E., and Weaver, W. E., *The Mathematical Theory of Communication*, University of Illinois Press, Urbana and Chicago, (1963).

[29] Cover, T. M., and Thomas, J. A., *Elements of Information Theory*, Wiley-Interscience, (1991).

[30] Resnikoff, H. L., *The Illusion of Reality*, Springer-Verlag, (1989).

[31] Pauli, W., *Pauli Lectures on Physics: Volume 5 - Wave Mechanics*, MIT Press, (1973).

[32] Goldin, E., *Waves and Photons: An Introduction to Quantum Optics*, John Wiley and Sons, (1982).

[33] Cohen-Tannoudji, C., *Quantum Mechanics*, Wiley, (1977).

[34] Ohya, M., *Some aspects of quantum information theory and their application to irreversible processes*, Reports on Mathematical Physics, Vol. 27, (1988).

[35] DiFrancia, T., *Resolving power and information*, Journal of the Optical Society of America, Vol. 45, No. 7, July 1955.

[36] Harris, J. L., *Diffraction and resolving power*, Journal of the Optical Society of America, Vol. 54, No. 7, July 1964.

[37] Gabor, D., Wolf, E. (ed), *Light and Information*, Progress in Optics1, 111-153, (1961).

{38] Underkoffler, J. S., *Personal communication.*

[39] Benton, S. A., *Experiments in holographic video imaging*, SPIE Institute Series, Vol. IS8, (1990).

[40] Benton, S. A., and Pappu, R. S., *Minimum-Pixel Holograms*, Proceedings of the IS&T/SPIE's Symposium on Electronic Imaging, Practical Holography IX, (1995).