Holographic Microscopy with a Complicated Reference

Abraham Szöke

Lawrence Livermore National Laboratory, Livermore, California 94550

In recent years three-dimensional images at atomic resolution have been obtained by holography as well as by x-ray crystallography. In this report we explore the connections between these two methods from a unified point of view. To recover the unknown structure we use mathematical methods developed for the solution of inverse problems. We review relevant experiments and discuss some ideas that may lead to more powerful imaging methods in the future.

Journal of Imaging Science and Technology 41: 332–341 (1997)

Introduction

When Gábor invented holography,[1,2] his goal was to obtain three-dimensional images at atomic resolution. His basic insight was that, if a known wavefront is made to interfere with an unknown wavefront, the phase of the unknown wavefront becomes measurable (with twofold uncertainty). He proposed a two-step method to recover three-dimensional objects using this phase information. In the first step, the recording of the hologram, the unknown object and a reference object are illuminated by a coherent source of waves. In its simplest version, the reference object is a small (pointlike) scatterer. The waves scattered from the reference object and those scattered from the unknown object form an interference pattern recorded on a screen at some distance away. Gábor called this recording a hologram, expressing that three-dimensional information is contained in the recorded interference pattern. The second step is the reconstruction of the scattering object from this recording. For this step he proposed to illuminate the hologram with a replica of the reference wave and showed that part of the wave transmitted by the hologram reproduces the wave originally scattered by the unknown object. Clearly a large and diverse range of applications has been built on Gábor's original ideas. Nevertheless, the central purpose of holography has stayed the same: to recover the object, the source of the object wave, from the intensities of the recorded interference pattern, the hologram.

Holograms at near-atomic resolution with a point reference were first produced by conventional, multikilovolt electron microscopes using beamsplitters (Mollenstedt biprisms) to produce a reference wave.[3,4] In-line Gábor holograms using soft x-rays have also been produced successfully.[5] A similar method that uses low-energy electrons was demonstrated recently by Morin et al.[6]: electrons, emitted by a single atom on a fine metal tip, produce holograms at potentially atomic resolution.

The most widely used and oldest known method for obtaining three-dimensional images at atomic resolution is x-ray crystallography. For years structures of very complex molecules have been unraveled by this method. The connection between x-ray (or electron, or neutron) diffraction methods and holography is intriguing. This paper attempts to illuminate that connection. We would like to stress that the connection has been well known for many years. In fact Gábor credits x-ray crystallographers with the fundamental ideas for holography. Specifically, the principles of holography will be extended to encompass x-ray diffraction. By doing so we hope that other scientific areas will also open for exploration.

To set the stage for the rest of the report, we discuss the fundamental and practical limitations of holographic microscopy: (1) The most fundamental limitations are determined by the wavelength, λ, of the incident wave and by the numerical aperture, NA, of the holographic system. As in all optical systems, the best transverse resolution obtainable is $\approx \lambda/(\mathrm{NA})$ and the longitudinal resolution is limited to $\approx \lambda/(\mathrm{NA})^2$. (2) The incident wave has to have enough coherence length in the appropriate directions that scattered waves from the reference and the object produce interference fringes. (3) If the reference point is at a macroscopic distance from the object, the recording medium itself has to have as good a resolution as the microscope to record the interference fringes. (4) The interference pattern between the reference wave and the object wave depends on the cosine of the phase difference between them; therefore, the intensity of the interference pattern can only give the absolute value of the phase and not its sign. This twofold ambiguity gives rise to the holographic dual image. When the coherence length of the source is very long, the reference can be sufficiently displaced from the unknown object so that the dual image is eliminated.[7] (5) In practice, electrical and mechanical stability requirements pose severe limitations on the resolution of holographic microscopy. (6) Biological specimens, whose imaging is of great interest, are damaged by the incident radiation, be it x-rays or electrons. To obtain an image, about 100 scattered particles have to be detected per voxel (three-dimensional resolution element).[†] For one-of-a-kind biological specimens, the damage limits the attainable resolution rather severely.[8,9] Fortunately, many interesting biological objects, for example proteins and RNA, come in large numbers of copies. If these molecules can be oriented in the same way (as they are in

Original manuscript received February 9, 1997.

crystals), the damage is reduced because the scattered and detected probe particles can come from any one of the identical copies of the specimen.

To circumvent some of the limitations mentioned above, the author has proposed to do holography with a local reference.[10] By local reference, we mean bringing the reference point close to the object and taking the detector screen far away. This procedure magnifies the interference pattern by the ratio of the two distances. The closeness of the reference point to the object eases coherence requirements of the source and the magnification relaxes demands on the resolution of the detector screen. If the specimen consists of many identical objects similarly oriented (e.g., macromolecular crystals), the holograms produced by each one of the copies are identical and can be reconstructed to give a single copy of the objects. This reduces the damage to the specimen, as discussed above. The arrangement can also alleviate stability problems and be made insensitive to lens aberrations. However, the arrangement also introduces some limitations. Both the reference and the unknown objects have to be microscopic and not too complicated. The dual image overlaps the desired image and is therefore difficult to eliminate.

As an example of holography with a local reference, the author of Ref. 10 points out that atoms that emit x-ray photons or electrons can serve as reference points. The waves emitted by those atoms are scattered by the surrounding medium. The interference between the unscattered wave and the scattered waves produces a hologram on a detector screen. Recently, Tegze and Faigel[11] and Gog et al.[12] have obtained x-ray holograms of crystals by variants of this method. Another practical method based on this principle uses photoelectrons emitted by atoms near a solid surface. This method is called photoelectron holography.[13,14]

The paper starts with a mathematical description of a hologram with a complicated local reference. Then we discuss the recovery of the unknown object from the hologram. We show that the recovery is analogous to well-known inverse problems. In the section **Holographic Interpretation of Some Experiments** we survey various experimental methods that can be perceived as holography with a complicated reference. Some are called holography, some are not but all are similar. In the short section **Proposed New Methods**, we explore some new experimental possibilities. These proposals are intended to show that a unified point of view can lead to new ideas. We close with a summary of what has been accomplished.

Holographic Equations for a Complicated Reference

Recovery of an unknown object from a hologram when the reference wave is a uniform plane or spherical wave consists of the following steps.[2,10] First, a hologram is recorded. Illumination of the hologram by a replica of the reference wave produces a replica of the object wave (among other waves). To recover the object, the wave is back propagated toward its source. Regions in space where the intensity of the back propagated wave is high are interpreted as the location of the unknown object. This procedure is analogous to the way luminous objects are seen by our eyes; three-dimensional resolution has the same limitations. We will now show how to do the analogous procedure to find the object from a hologram with a complicated reference.

Let the vector $\mathbf{R}$ denote points on the (spherical) holographic screen where the interference pattern is recorded. Similarly, $\mathbf{x}$ will denote the coordinates in the three-dimensional space where the reference and the object are located. Let $R(\mathbf{R})$ denote the complex amplitude of an (arbitrarily complicated) reference wave on the hologram screen. Assume that a simple object of unit strength is located at a single point, $\mathbf{x}_n$ and is illuminated by an incident wave. It could be, e.g., a point scatterer or a Gaussian blob of extent Δx. Its density will be denoted by $\Omega(\mathbf{x}-\mathbf{x}_n)$. The object wave on the screen, $O_n(\mathbf{R})$, which is elastically scattered by the simple object, can be calculated from scattering theory.[15] In the first Born approximation it is

$$O_n(\mathbf{R}) = \psi_O(\mathbf{R}) + \int G(\mathbf{R}-\mathbf{x})\,\tau(\hat{\mathbf{R}},\hat{\mathbf{k}}_{in},\mathbf{x}-\mathbf{x}_n)\,\psi_O(\mathbf{x})d\mathbf{x}, \quad (1)$$

where $\psi_O(\mathbf{R})$ and $\psi_O(\mathbf{x})$ are the incident (scalar) wave amplitudes at the recording screen and at the scatterer respectively. The unit vectors $\hat{\mathbf{k}}_{in}$ and $\hat{\mathbf{R}}$ denote the directions of the incident and scattered fields, respectively, and $\tau(\hat{\mathbf{R}},\hat{\mathbf{k}}_{in},\mathbf{x}-\mathbf{x}_n)$ denotes the angle- and space-dependent elastic scattering amplitude of the simple scatterer of unit strength $\Omega(\mathbf{x}-\mathbf{x}_n)$ centered on $\mathbf{x}_n$ (also known as the t matrix.) The propagator, $G(\mathbf{R}-\mathbf{x})$, which describes the amplitude of the field at $\mathbf{R}$ that originates at $\mathbf{x}$,

$$G(\mathbf{R}-\mathbf{x}) = -\frac{\exp\left(ik|\mathbf{R}-\mathbf{x}|\right)}{4\pi|\mathbf{R}-\mathbf{x}|}, \quad (2)$$

where k is the wave vector of the propagating wave field. In the applications envisaged, the distance to the screen is much larger than the size of the object. Therefore the far-field approximation suffices:

$$G(\mathbf{R}-\mathbf{x}) \approx -\frac{\exp\left(ik|\mathbf{R}|\right)}{4\pi|\mathbf{R}|}\exp(i\mathbf{k}_{out}.x), \quad (3)$$

where $\mathbf{k}_{out} = k\hat{\mathbf{R}} = k\mathbf{R}/|\mathbf{R}|$ is the wave vector of the scattered wave in the direction of the observation. The terms neglected are of relative magnitude $|\mathbf{x}|/|\mathbf{R}|$. A fast algorithm for the calculation of $O_n(\mathbf{R})$ to all orders has recently been derived.[16]

When the incident wave is a plane wave, it is described by

$$\psi_O(\mathbf{x}) = \Psi_O\exp(i\mathbf{k}_{\text{in}}\cdot\mathbf{x}). \quad (4)$$

If the incident wave $\psi_O(\mathbf{R})$ is filtered out before it gets to the screen, the wave amplitude on the screen can be written as

$$O_n(R) = -\int \frac{\exp\left(ik|R|\right)}{4\pi|\mathbf{R}|} \tau(\hat{\mathbf{R}},\hat{\mathbf{k}}_{in},\mathbf{x}-\mathbf{x}_n)\Psi_O\exp\left[-i\left(\mathbf{k}_{out}-\mathbf{k}_{in}\right)\cdot\mathbf{x}\right]d\mathbf{x}. \quad (5)$$

If the screen is spherical, making $|\mathbf{R}|$ constant, the object wave depends only on the momentum transfer vector $\mathbf{h} = (\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}})/2\pi$. Thus Eq. 5 can be written symbolically as

$$O_n(\mathbf{h}) = \int\rho_n(\mathbf{x})\exp(-2\pi i\mathbf{h}\cdot\mathbf{x}), \quad (6)$$

where scale factors have been absorbed into the definition of $\rho_n(\mathbf{x})$. Note that apart from scale factors and complex conjugation this is the notation used in x-ray crystallography.

Assuming that the reference and object waves are coherent, the intensity of the recorded hologram can, of course, be calculated:

$$I_n(\mathbf{h}) = |R(\mathbf{h}) + O_n(\mathbf{h})|^2. \quad (7)$$

Turning now to the general case of an arbitrary object, let $\rho(\mathbf{x})$ denote the density of the unknown object and $\tau(\hat{\mathbf{R}}, \hat{\mathbf{k}}_{in}, \mathbf{x})$ its scattering power. We will denote by $\rho_{app}(\mathbf{x})$ an approximation to $\rho(\mathbf{x})$, obtained by decomposing it into a suitable set of basis functions $\Omega(\mathbf{x} - \mathbf{x}_n)$ of scattering power, $\tau(\hat{\mathbf{R}}, \hat{\mathbf{k}}_{in}, \mathbf{x} - \mathbf{x}_n)$, (see Appendix). The expansion is given in terms of the unknown quantities N_n that measure the unknown strength of the scatterer at the vicinity of $\mathbf{x}_n$. It is

$$\rho_{\text{app}}(\mathbf{x}) = \sum_n N_n \Omega(\mathbf{x} - \mathbf{x}_n), \quad (8)$$

giving the diffracted amplitude

$$O(\mathbf{h}) = \sum_n N_n O_n(\mathbf{h}), \quad (9)$$

where $O_n(\mathbf{h})$ is defined in Eq. 6. The hologram intensity is, therefore, given by

$$\begin{aligned} I(\mathbf{h}) &= \left|R(\mathbf{h}) + O(\mathbf{h})\right|^2 = \\ &= \left|R(\mathbf{h})\right|^2 + \sum_n N_n \left[O_n^*(\mathbf{h}) R(\mathbf{h}) + O_n(\mathbf{h}) R^*(\mathbf{h})\right] + \\ &+ \left|\sum_n N_n O_n(\mathbf{h})\right|^2. \end{aligned} \quad (10)$$

To repeat, we derived Eq. 10 under the assumption of an incident plane wave and we used the far-field approximation. Both these conditions are satisfied if the sizes of the reference object and the unknown object as well as their mutual distance are much smaller than their distances to the source of incident radiation and to the recording screen.

The recovery of the unknown object has thus been translated in Eq. 10 into the solution of a set of inhomogeneous quadratic equations in the unknown quantities N_n. The significance of this simple derivation[17] is that it shows the complete analogy of holography with other inverse problems.[18–21] Like the analogous equations in other inverse problems, Eq. 10 is expected to be ill-conditioned. A (nonlinear) holographic operator can be defined as one that produces the hologram $I(\mathbf{h})$ from the set of scatterers N_n and from the reference wave. The technical definition of ill conditioning is that the ratio of the largest to the smallest singular value of this holographic operator is much larger than unity. This means the solution of Eq. 10 is very poorly defined and is also extremely sensitive to noise in the data. Note that this property of inverse problems is inherent and does not depend on our particular formulation. The advantage of recognizing the analogy of holography and inverse problems is that the vast knowledge gathered in the solutions of inverse problems can be applied directly to holography.

Holographic Dual Image

Further light can be shed on our method by returning to the (idealized) analysis of the traditional recovery method of holography and its limitations.[2] The intensity of the recorded hologram is given by Eq. 10. For the present discussion a strong reference beam will be assumed and the self-interference of the object wave will therefore be neglected. This results in a net hologram intensity,

$$H(\mathbf{h}) = I(\mathbf{h}) - |R(\mathbf{h})|^2 \approx R^*(\mathbf{h}) O(\mathbf{h}) + R(\mathbf{h}) O^*(\mathbf{h}) \equiv \mathcal{L}(N_p). \quad (11)$$

The last equality means that, in this approximation, $H(\mathbf{h}) = \mathcal{L}(N_p)$ is a linear operator operating on the finite dimensional vector of discrete amplitudes. In the second step of Gábor's reconstruction, the hologram is illuminated with a replica of the reference wave. The (idealized) wave transmitted through the hologram is the reference wave modulated by the hologram. It is obtained by multiplying Eq. 11 by $R(\mathbf{h})$,

$$R(\mathbf{h})H(\mathbf{h}) = |R(\mathbf{h})|^2 O(\mathbf{h}) + R(\mathbf{h})^2 O^*(\mathbf{h}). \quad (12)$$

Suppose now that the reference wave is produced by a single point scatterer located at position $\mathbf{x}_O$. According to Eq. 5 this produces a reference wave proportional to $\exp(-2\pi i \mathbf{h} \cdot \mathbf{x}_O)$. It can be written as

$$R(\mathbf{h}) = R_O \exp(-2\pi i \mathbf{h} \cdot \mathbf{x}_O). \quad (13)$$

Dividing Eq. 12 by $2\,|R(\mathbf{h})|^2$ from Eq. 13 results in

$$\frac{R(\mathbf{h})H(\mathbf{h})}{2{R_O}^2} = \frac{1}{2}\left[O(\mathbf{h}) + O^*(\mathbf{h}) \exp(-4\pi i \mathbf{h} \cdot \mathbf{x}_O)\right]. \quad (14)$$

Substituting from Eqs. 6, 8, and 9 for the unknown scatterer density and assuming that it is real, we have

$$\begin{aligned} &\frac{R(\mathbf{h})H(\mathbf{h})}{2R_O^2} = \\ &\frac{1}{2}\left[\int \rho_{app}(\mathbf{x}) \exp\{-2\pi i \mathbf{h} \cdot \mathbf{x}\} d\mathbf{x} + \int \rho_{app}(\mathbf{x}) \exp\{-2\pi i \mathbf{h} \cdot (2\mathbf{x}_O - \mathbf{x})\} d\mathbf{x}\right]. \end{aligned} \quad (15)$$

The second integral can be transformed by the substitution $2\mathbf{x}_O - \mathbf{x} \to \mathbf{x}$ giving

$$\begin{aligned} &\frac{R(\mathbf{h})H(\mathbf{h})}{2{R_O}^2} = \\ &\frac{1}{2}\left[\int \rho_{\text{app}}(\mathbf{x}) \exp\{-2\pi i \mathbf{h} \cdot \mathbf{x}\} d\mathbf{x} + \right. \\ &\left. \int \rho_{\text{app}}(2\mathbf{x}_O - \mathbf{x}) \exp\{-2\pi i \mathbf{h} \cdot \mathbf{x}\} d\mathbf{x}\right]. \end{aligned} \quad (16)$$

This equation can be solved by discrete inverse Fourier transformation if and only if the conditions of the sampling theorem (Nyquist conditions) are satisfied. The result is the sum of two densities symmetrically located around the reference point $\mathbf{x}_O$. The reference point, in fact, introduces a center of symmetry. In holography this image doubling has been known ever since Gábor's original papers.

The image obtained by Gábor's reconstruction is not the most general holographic image, i.e., it is not the most general solution of Eq. 11. In fact, any linear superposition of the two images $\rho(\mathbf{x}) = (1 - \mu)\, \rho_{\text{app}}(\mathbf{x}) + \mu\, \rho_{\text{app}}(2\mathbf{x}_O - \mathbf{x})$ with arbitrary real μ satisfies Eq. 10. In holography using

laser light, the reference point is usually shifted so far away that the dual image does not overlap the real one.[7] If known, the spatial extent of the original image can be used as a constraint in the solution of Eq. 11. This results in setting $\mu \to 0$, thereby getting the correct image, uncontaminated by its dual.

However, in holograms with a complicated reference, the reference wave $R(\mathbf{h}) = |R(\mathbf{h})| \exp[i\varphi(\mathbf{h})]$ is itself complicated. Also, the two images may overlap to a large extent. To analyze this situation, Gábor's procedure, valid for a single point scatterer, will now be generalized. Equation 12 is formally divided by $2|R(\mathbf{h})|^2$, giving

$$\frac{R(\mathbf{h})H(\mathbf{h})}{2|R(\mathbf{h})|^2} = \frac{1}{2}\left[O(\mathbf{h}) + O^*(\mathbf{h})\frac{R(\mathbf{h})^2}{|R(\mathbf{h})|^2}\right]. \tag{17}$$

The term $R(\mathbf{h})^2/|R(\mathbf{h})|^2 = \exp[i2\varphi(\mathbf{h})]$ is a pure phase term that is never singular. The term on the left side becomes indeterminate for a given $\mathbf{h}$ if $|R(\mathbf{h})|^2 \to 0$. In the mathematical sense, $H(\mathbf{h})$ vanishes at least linearly for the same $\mathbf{h}$, but numerically the value of the left hand side of Eq. 17 is very sensitive to inaccuracies in the measured value of $I(\mathbf{h})$. This is a manifestation, in $\mathbf{h}$ space, of the ill conditioning of Eq. 9 or, equivalently, the poor phasing power or quality of the particular Fourier component $R(\mathbf{h})$ of the reference.

The two terms on the right side will now be viewed as the diffraction pattern of the real image and of the (holographic) dual image. From Eqs. 6 and 9

$$O(\mathbf{h}) = \int \rho_{\text{app}}(\mathbf{x}) \exp\{-2\pi i \mathbf{h} \cdot \mathbf{x}\} d\mathbf{x}. \tag{18}$$

If $O(-\mathbf{h}) = O^*(\mathbf{h})$, it follows that $\rho_{\text{app}}(\mathbf{x})$ is real. The dual density is defined implicitly by

$$O^*(\mathbf{h}) \exp[i2\varphi(\mathbf{h})] = \int \rho_{\text{dual}}(\mathbf{x}) \exp\{-2\pi i \mathbf{h} \cdot \mathbf{x}\} d\mathbf{x}, \tag{19}$$

and again, if $O(-\mathbf{h}) = O^*(\mathbf{h})$, $\rho_{\text{dual}}(\mathbf{x})$ is real.

The dual image has several interesting properties: it is a linear function of $\rho_{\text{app}}(\mathbf{x})$, but a nonlinear function of $R(\mathbf{h})$. For a complicated reference the positivity of $\rho_{\text{app}}(\mathbf{x})$ does not imply positivity of $\rho_{\text{dual}}(\mathbf{x})$ and if $\rho_{\text{app}}(\mathbf{x})$ has a known, limited support[‡] $\rho_{\text{dual}}(\mathbf{x})$ does not necessarily have the same support. The hologram of the dual image is the same as that of the correct image; in particular, it follows (from the $\mathbf{h} = 0$ component) that the two images have the same total scattering strength. Verification that our definition of the dual image tends to its correct limits is easy: When the reference is a single point, it is the image centrally inverted with respect to that point (Eq. 16) and when the reference disappears, it is the enantiomorph[§] of the object. Note that all the preceding properties of the dual image are unchanged when the $|O|^2$ term is included.

In analogy to holography with a simple reference, Gábor's reconstruction method does not produce the most general solution of Eq. 11. The back transform of Eq. 17 always gives the superposition $(1/2)[\rho_{\text{app}}(\mathbf{x}_p) + \rho_{\text{dual}}(\mathbf{x}_p)]$. To establish the connection between the formulation using linear algebra and the holographic analogy, we observe that

$$\mathcal{R}\text{e}\{R^*(\mathbf{h})O(\mathbf{h}) - R(\mathbf{h})O^*(\mathbf{h})\} = 0. \tag{20}$$

Therefore the addition of $\rho_{\text{app}}(\mathbf{x}_p) - \rho_{\text{dual}}(\mathbf{x}_p)$, to the reconstructed image cannot change the discrepancy between the two sides of Eq. 11. In addition, the linear operator $\mathcal{L}(N_p)$ operating on $\rho_{\text{app}}(\mathbf{x}_p) - \rho_{\text{dual}}(\mathbf{x}_p)$ gives zero as well. This proves $\mathcal{L}(N_p)$ has at least one null vector.

The most general solution of Eq. 11 is well known from linear algebra to contain nonzero singular vectors of the linear operator of Eq. 11, $\mathcal{L}(N_p)$, augmented with an arbitrary vector from its null space. For the general treatment we refer to Golub and Van Loan.[22] The number of independent null vectors of Eq. 11 will be denoted by N_{null}. We have shown above that $N_{\text{null}} \geq 1$.

Denoting the orthogonal basis of right singular vectors of the null space by $\mathbf{v}_{\text{null,j}}$ the most general real solution of Eq. 11 is given by

$$\rho(\mathbf{x}_p) = \frac{1}{2}\left[\rho_{\text{app}}(\mathbf{x}_p) + \rho_{\text{dual}}(\mathbf{x}_p)\right] + \sum_{\text{j}=1}^{N_{\text{null}}} \mu_{\text{j}} \mathbf{v}_{\text{null,j}}, \tag{21}$$

where μ_j are a set of arbitrary real numbers. The particular (Gábor) solution, $[\rho_{\text{app}}(\mathbf{x}_p) + \rho_{\text{dual}}(\mathbf{x}_p)]/2$, has a conjugate null image $[\rho_{\text{app}}(\mathbf{x}_p) - \rho_{\text{dual}}(\mathbf{x}_p)]/2$. Singling out this null vector and orthogonalizing the rest of the null space to it, we see, after a small amount of algebra, that the most general solution of Eq. 11 may be written equivalently as

$$\rho(\mathbf{x}_p) = (1-\mu)\rho_{\text{app}}(\mathbf{x}_p) + \mu\rho_{\text{dual}}(\mathbf{x}_p) + \sum_{\text{j}=1}^{N_{\text{null}}-1} \mu_{\text{j}} \mathbf{v}_{\text{null,j}}, \tag{22}$$

where μ, μ_j are a set of arbitrary real numbers. In this notation the desired solution is $\mu = \mu_j = 0$, while the minimum norm least squares solution is $\mu = 1/2$, $\mu_j = 0$. Note that, from the point of view of information theory, N_{null} pieces of information are still missing. The formulation presented here allows the easy incorporation of additional information that restricts the solution space. A similar derivation can be carried out for the full quadratic equation, Eq. 12.

Methods of Solution

We stated in **Holographic Equations for a Complicated Reference** above, that holographic recovery is in essence an inverse problem, and as such, is almost always ill conditioned.[18–21] In this section we elaborate on the uniqueness of the solution and discuss some of the properties of the algorithms used for recovery of the scatterer. The most fundamental requirement for a unique solution of the recovery problem is that enough information be available from experiments. Many years ago the eminent mathematician Lánczos[23] pointed out that no mathematical trickery can remedy lack of information.

The maximum amount of available information is significantly limited in diffraction experiments on periodic structures. When waves diffract from crystals, the diffraction spots satisfy Bragg's conditions. In more technical language, if the crystal consists of well-ordered N unit cells and if the kinematic (first-order Born) approximation holds,[24] the reflection intensity is concentrated into a volume of reciprocal space of $\approx 1/N$ and has a peak intensity $\approx N^2$ that corresponds to an integrated intensity $\approx N$. Both these properties are important for experiments on complicated molecules that otherwise would be destroyed by incident x-rays.[8,9] The diffraction pattern is characterized by the momentum transfer vector $\mathbf{h} = (\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}})/2\pi$, which corresponds to a resolution of $1/|\mathbf{h}|$. If data are available up to a given resolution, the details of the scatterer can be reconstructed only to that resolution.**

Counting the number of diffraction spots to a given resolution is easy. If the phases of the Bragg reflections are known, the electron density can be reconstructed uniquely from the diffraction pattern, and the number of reflections within a resolution shell correspond exactly to the critical Nyquist sampling frequency.[25] In diffraction experiments only the amplitudes are measurable. Let us assume that the number of bits of information supplied by the amplitudes of the structure factors is equal to that supplied by their phases. Then absence of phases corresponds to the loss of exactly one half of the information. In crystal diffraction, if we write the Bragg conditions in the usual way as $d = \lambda/(2\sin\theta)$, we can see that changing the x-ray wavelength changes only the diffraction angle that corresponds to a given d in the crystal. Therefore, if the fundamental scattering amplitude $\tau\left(\hat{\mathbf{R}}, \hat{\mathbf{k}}_{\text{in}}, \mathbf{x} - \mathbf{x}_n\right)$ in Eq. 1 is independent of the x-ray energy and the scattering angle, we get no new information from experiments at different wavelengths. In x-ray crystallography additional information is obtained by anomalous scattering, i.e., the dependence of the scattering amplitude on wavelength, by adding heavy atoms to the crystal, or by having prior knowledge of the electron density of certain parts of the crystal. For example, usually about half the volume of the unit cell of proteins contains disordered water with nearly uniform density.

The situation is radically different in photoelectron holography in its many variations and in x-ray (emission) holography. First, the full diffraction pattern is observable, so it can be sampled at the Nyquist frequency or even more finely. Second, by doing the experiment at several different energies we obtain additional information. Third, in electron holography the scattering amplitudes of low-energy electrons usually depend strongly on energy and angle.

Barton's first method for recovering the scatterers in photoelectron holography used Gábor's algorithm.[27] Barton's method was essentially the solution of Eq. 15 by an inverse Fourier transformation. The interpretation of the result is that the scatterers are located where the value of $\rho(\mathbf{x})$ is large. The recovered density suffered from various shifts in positions and splitting as a consequence of the angular structure of the complex t matrix, $\tau\left(\hat{\mathbf{R}}, \hat{\mathbf{k}}_{\text{in}}, \mathbf{x}\right)$ of the scatterers. Several ways exist to correct for this difficulty; if data are collected at several different photoelectron energies, a phased addition of the holograms reinforces the real image and tends to destroy the dual image.[28,29] The basis function expansion of this paper was also generalized for the recovery of scatterers with known complex angle-dependent and energy-dependent scattering amplitudes.[30,31]

Holographic Interpretation of Some Experiments

X-Ray Crystallography. To show the connection between x-ray crystallography and holography,[15] we divide the electron density of the crystal, perhaps artificially, into a known and an unknown part: $\rho(\mathbf{r}) = \rho_{\text{kno}}(\mathbf{r}) + \rho_{\text{unk}}(\mathbf{r})$. In a perfect crystal, the electron density is a periodic function. The scattering amplitude in the x-ray regime is proportional to the electron density and the incident x-rays are very nearly plane waves. The Fourier transform of the electron density $F(\mathbf{h})$ is called the *structure factor* in crystallography. Structure factors can be defined similarly for the known and the unknown part as follows:

$$R(\mathbf{h}) = \int_{\text{unit cell}} \rho_{\text{kno}}(\mathbf{r}) \exp(2\pi i \mathbf{h} \cdot \mathcal{F}\mathbf{r}) d\mathbf{r}, \tag{23}$$

$$O(\mathbf{h}) = \int_{\text{unit cell}} \rho_{\text{unk}}(\mathbf{r}) \exp(2\pi i \mathbf{h} \cdot \mathcal{F}\mathbf{r}) d\mathbf{r}, \tag{24}$$

where $\mathcal{F}$ denotes the transformation of $\mathbf{r}$ into fractional coordinates of the unit cell of a crystal[17] and $\mathbf{h}$ denotes a triplet of integers pointing to the lattice points in reciprocal space. The measured x-ray diffraction intensities can be reduced to the square magnitude of the structure factor of the crystal $F(\mathbf{h})$ producing the formula

$$\begin{aligned} |F(\mathbf{h})|^2 &= |R(\mathbf{h}) + O(\mathbf{h})|^2 = \\ &= |R(\mathbf{h})|^2 + R(\mathbf{h})O^*(\mathbf{h}) + R^*(\mathbf{h})O(\mathbf{h}) + |O(\mathbf{h})|^2. \end{aligned} \tag{25}$$

Comparison of Eq. 25 with Eq. 10 shows clearly the analogy with holography. Also clear is that Eq. 10 can be used to find the unknown part of the molecule and that completion of a crystal structure, when a part of the unit cell has already been found, is similar to many other inverse problems. Equation 25 highlights that when a large part of the electron density is known, the interference terms between the reference wave and the object wave give most of the information on the missing part of the structure.

The approach outlined above has been developed extensively by the author and collaborators into a software package *EDEN* (for Electron DENsity) for macromolecular x-ray crystallography. In the following, we discuss some of the detailed derivations of the algorithm used in EDEN. The unknown part of the electron density is described as a sum of Gaussian basis functions of equal widths, centered on a grid that divides the unit cell into P_a, P_b, and P_c equal parts along the crystallographic axes **a**, **b**, and **c**, respectively. The grid points are denoted by $\mathbf{r}_p$, $p = 1, \ldots, P$ where $P = P_a \cdot P_b \cdot P_c$. Each Gaussian blob (voxel) is assumed to contain an unknown number of electrons, $n(p)$:

$$\rho_{\text{unk}}(\mathbf{r}) \approx \frac{1}{(\pi\eta\Delta r^2)^{3/2}} \sum_{p=1}^{P} n(p) \exp\left[\frac{-\left|\mathbf{r} - \mathbf{r}_p\right|^2}{\eta\Delta\mathbf{r}^2}\right], \tag{26}$$

where Δr is the mean grid spacing and η determines the width of the Gaussians relative to the grid spacing. When Eq. 26 is extended periodically over the repetitions of the unit cell, a simple derivation results in the following formula for the structure factors of the unknown part $O(\mathbf{h})$:

$$O(\mathbf{h}) = \exp\left[-\eta\left(\pi\Delta r\left|\mathcal{F}^T\mathbf{h}\right|\right)^2\right] \sum_{p=1}^{P} n(p) \exp(2\pi i \mathbf{h} \cdot \mathcal{F}\mathbf{r}_p). \tag{27}$$

When the representation of the unknown density is substituted from Eq. 27, Eq. 25 becomes a set of quadratic equations in the unknowns $n(p)$. The number of equations N_h is usually not equal to the number of unknowns P. The equations may contain inconsistent information for many reasons. Examples include experimental errors, changes in the molecular structure when heavy atoms are added to the molecule, imperfect knowledge of molecular fragments, and incomplete noncrystallographic symmetry. The equations are ill-conditioned, and therefore their solutions are extremely sensitive to noise in the data. For these conditions, the equations may have many solutions or no solution at all. Our way of circumventing these problems is to obtain a quasi-solution of Eq. 25 by minimizing the discrepancy, or cost function (see e.g., Dainty and Fienup[32]):

$$f_{\text{eden}} = \frac{1}{2} \sum_{\mathbf{h}} w'(\mathbf{h})^2 \left[\left|R'(\mathbf{h}) + O(\mathbf{h})\right| - \left|F'(\mathbf{h})\right|\right]^2, \tag{28}$$

where $R'(\mathbf{h})$ and $F'(\mathbf{h})$ are apodized or smeared versions of $R(\mathbf{h})$ and $F(\mathbf{h})$ and where $w'(\mathbf{h})^2$ is a set of positive weights, usually set to unity or to $1/\sigma^2(\mathbf{h})$, the inverses of the variances of the measured Bragg reflection intensities.

The effective or intrinsic resolution of the observed structure factor amplitudes is determined by atomic structure factors, atomic motion, and crystalline disorder. The intrinsic resolution of the Gaussian basis set (Eq. 26) is $\eta\Delta r^2$. If the effective resolution of the measured structure factor amplitudes is higher than that of the Gaussian basis functions, they have to be modified to

$$|F'(\mathbf{h})| = |F(\mathbf{h})| \exp\left[-\delta\eta\left(\pi\Delta\mathbf{r}\left|\mathcal{F}^T\mathbf{h}\right|\right)^2\right], \qquad (29)$$

using a nonnegative parameter δ. This procedure, usually called apodization, adjusts the resolution of the measured diffraction pattern to that of the Gaussian basis set used in the solution. Apodization is equivalent to an appropriate smearing of the electron density of the molecule by a Gaussian smearing function. The smearing is essential for a mathematically stable fitting of the high-resolution reflections. An analogous procedure is used to adjust the intrinsic resolution of the known part. The summation in Eq. 28 includes only available experimental data; values of $R(\mathbf{h})$ for which the corresponding $F(\mathbf{h})$ is missing are left indeterminate by setting $w'(\mathbf{h})^2$ to zero. Therefore truncation errors of Fourier inversions are absent and the consistent use of nonnegative basis functions is fully justified.

A second type of information on the crystal structure is a partial knowledge of the electron density in parts of the unit cell. For example, part of the molecule may be very similar to another molecule whose structure is known. As another example, the solvent volume has a featureless electron density of a well-known value. Such knowledge can be incorporated into EDEN as a target density, expressed in terms of the amplitudes of the basis functions used in the main program. They will be denoted by $n(p)_{\text{target}}$. There is a corresponding cost function

$$f_{\text{space}} = \frac{1}{2}\lambda_{\text{space}} P \sum_{p=1}^{P} \tilde{w}_p{}^2 \left\{n(p) - n(p)_{\text{target}}\right\}^2. \qquad (30)$$

The overall relative weight λ_{space} and the individual weights at each point $\tilde{w}_p{}^2 \le 1$ express the strength of our belief in the correctness of the target density: the weights $\tilde{w}_p{}^2$ may be used to emphasize or deemphasize different regions of the target density (although usually they are set to 1 or 0), while λ_{space} determines the relative importance of f_{space} with respect to f_{eden}. In the presence of a target density, the actual cost function used in the computer program is the sum of f_{eden} (Eq. 28) and f_{space} (Eq. 30):

$$f_{\text{total}} = f_{\text{eden}} + f_{\text{space}}. \qquad (31)$$

Additional information leads to additional terms in the cost function (Eq. 31).

As seen from Eq. 27 the structure factors of the unknown part can be calculated by fast Fourier transforms followed by a scalar multiplication. The gradients of the cost function can be calculated similarly. This leads to a fast (P-log P) algorithm and to the ability to minimize Eq. 31 without ever calculating or saving $P \times P$ matrices. In EDEN the cost function is minimized using a conjugate gradient algorithm that is very efficient in the presence of nonlinear constraints. A basic constraint of nonnegativity of the electron density is incorporated directly into the conjugate gradient optimizer by stipulating that all elements of the solution vector $n(p)$, be nonnegative.

The representation of the unknown density Eq. 26 uses an overcomplete set of Gaussian basis functions not orthogonal to each other. The usefulness and accuracy of the method are determined by the answers to the following mathematical questions: How well can the electron density of an arbitrary molecule be approximated by the superposition of such basis functions (with appropriate coefficients)? Is there a well-defined algorithm to find such a set of coefficients, given the electron density? Is the set of coefficients unique? Finally, if two sets of coefficients are similar (close to each other), are the resulting electron densities close to each other? Fortunately, mathematicians have done extensive research on such nonorthogonal, redundant basis sets called frames. Excellent discussions can be found in a book by Daubechies[33] and in a review by Heil and Walnut.[34] Some of their important results are described in the **Appendix** and are summarized below.

The mathematicians' answer to the first question is that electron densities can indeed be approximated well by such representations, if the electron density does not vary too wildly. Restated in technical language, the requirements are that the diffraction pattern and the basis set have similar intrinsic resolutions and that the grid spacing be about twice as fine as required by the corresponding Nyquist criterion. In our algorithm, this is achieved by the appropriate choices of η and Δr in Eq. 26. Also true is that two representations with similar coefficients do yield similar electron densities. Conversely, two similar electron densities produce similar sets of coefficients in our algorithm, which is therefore mathematically stable. But a given electron density can be represented by several different sets of coefficients. In fact, there are many possible algorithms to find a set of coefficients that approximate the electron density of the crystal equally well, of which the algorithm used by EDEN is only one example.

The holographic method has several advantages over other techniques in use today. Holography can incorporate external information clearly and consistently. For example, the positivity of the electron density, the near uniformity of the solvent region, or the similarity of parts of one molecule to parts of another are very important constraints on the electron density and, therefore, lead to a successful solution of macromolecular structures. In crystallographic jargon, EDEN is capable of changing the phases of calculated structure factors. Furthermore, EDEN lends itself to accurate mathematical analysis and it can be implemented on the computer using fast Fourier transforms. This enabled us to produce a computer program that solves large crystallographic problems of current interest in reasonable computer time. Incorporating the positivity constraint bridges the gap to some extent, that existed between direct and other methods in crystallography. The holographic method also is optimal in that, computationally its solution degrades gracefully with added noise. Our experience with the method has been published in a series of papers in *Acta Crystallographica*.[17,35–38] Those papers also present examples of protein crystals recovered by EDEN.

Electron Microscopy of Two-Dimensional Crystals. Some large molecules, especially membrane proteins, have been crystallized in two-dimensional form. These crystals have been examined by electron microscopes both in a focused and in a diffractive mode.[39–41] The image obtained by

the focused electron beam has been used to determine the phases of the low-resolution diffraction spots obtained by the diffraction of the unfocused electron beam. The subsequent processing of the image has been done in analogy to x-ray diffraction. The result has been the recovery of the three-dimensional structure of the protein. It is clear from the preceding that if part of the structure is either known or can be guessed, recovery of the rest of the molecule can again be viewed as a holographic inverse problem. We propose that the application of holographic methods may improve the quality of the recovery in the future. In particular, the difference in transverse and longitudinal resolution of the recovery may be reduced.

Low Energy Electron Diffraction. The diffraction of low-energy electrons from a crystal surface is closely related to the diffraction of x-rays by crystals, and therefore we expect that similar ideas should be applicable.[30] Low-energy electron diffraction (LEED) is also capable of atomic resolution. The scattering of such (100 to 1000 eV) electrons is much stronger than that of x-rays. This makes LEED surface-sensitive, but makes the analysis of the diffraction pattern more difficult, because multiple scattering of the diffracted electrons can be very important. A relatively simple case to analyze is a crystal surface with a submonolayer of adsorbed atoms in equivalent positions. The resulting diffraction pattern is similar to a conventional hologram.[42] If the adsorbate is a molecule, the observed electron diffraction pattern can still be interpreted as a hologram with a complicated reference. The equations to be solved are analogous to Eq. 10 above, except for the difficulty with multiple scattering. Nevertheless, we expect the method of Eq. 10 will extend the reach of low-electron energy diffraction to more complex overlayers. For a detailed review of this topic we refer to Refs. 13 and 14.

Other Related Methods. In a recent experiment, Tegze and Faigel[11] invoked holography with a local reference explicitly. They irradiated a $SrTiO_3$ crystal by x-rays of 17.4 keV with photon energy above the *K*-edge of Sr and carefully measured the angular distribution of the characteristic *K* radiation emitted by the Sr atoms. The fluorescence that arrived at the detector without scattering was the reference wave and the radiation scattered by the neighboring atoms was the object wave. This was therefore a holographic experiment with a simple reference.[10] The authors reconstructed the neighboring Sr atoms successfully, using a holographic recovery algorithm. In another new development, Gog et al.[12] extended holography with a local point reference to a "time-reversed" experiment. They irradiated an Fe_2O_3 crystal above the *K*-edge of iron and monitored the Fe fluorescence intensity as they changed the direction of the incident beam. The interference of the incident unscattered wave with the wave scattered by the neighboring atoms results in an angular dependence of the Fe fluorescence. The angular pattern of the fluorescence was successfully solved to yield the positions of the neighboring iron atoms in three dimensions. One advantage of the method is that variation of the incident x-ray wavelength gives additional information, as discussed in **Methods of Solution**. In both these experiments the contrast of the fringes is very weak and therefore restricted to damage-resistant materials.[8,9]

Electron holography using low-energy electrons was carried out by Morin, Pitaval, and Vicario.[6] The electrons were generated by field emission from a single atom of a fine tungsten tip. The beam was split using a biprism. One part of the beam produced the reference and the other part went through the object. The result was a standard off-axis hologram, that was successfully solved to yield the object, a set of carbon fibers at 7-nm resolution. Although the method does not automatically solve the stability problems of electron holography, the technique is a very promising development because low-energy electrons apparently cause relatively little damage in biological materials.[43]

The method of Rodenburg et al.[44–46] in electron microscopy can be perceived as holography without a reference wave or holography where a complicated reference is reconstructed simultaneously with the unknown object. The essence of the method is to focus the electron microscope on a (planar) sample and to look at the interference of two diffraction orders, made to overlap at the detector. The interference contains holographic information. The focal point of the incident beam is scanned over the sample, and the unknown object is reconstructed using Wigner functions. The information obtained is similar to the autocorrelation function of the object (the so-called Patterson function in crystallography). Clearly the position of the incident beam has to be very stable. Moreover, the method does not alleviate the damage problem in biological specimens.

We would like to point out that x-ray microscopy is also a related subject. Some excellent reviews of its application to biological specimens have appeared recently.[8,47]

Proposed New Methods

We now propose two new methods for holographic microscopy with a complicated reference. These sample ideas are presented as illustrations to the many possibilities that arise from our considerations.

For an example of an experiment possible with a low-energy electron beam, let us assume a large molecule is laid down on a substrate. An example of such a substrate is a Si crystal patterned on an appropriate scale. An example of a molecule is a short strand of DNA. Such a molecule has been found quite stable in a low-energy electron beam.[6,15,43] The angular distribution of the scattered electron beam could easily be detected by standard LEED techniques. The distribution consists of a discrete diffraction pattern due to the substrate and a diffuse pattern due to the adsorbed molecule. If the crystal is good enough and cold enough, it will not contribute appreciably to the diffuse pattern. The diffuse pattern could be analyzed using the theoretical techniques developed in this report. In particular, if part of the molecule is known, the rest could be found. The atoms of the substrate crystal may also be used as part of the reference. This proposal is similar to that of Ref. 48. An important point to keep in mind is that the method is sensitive only to the local order. One of the difficulties of such an experiment is that the penetration depth of low-energy electrons is very short and the sample has to be kept in vacuum.

A different kind of holography with a complicated reference could be done in protein crystals. The amino acid methionine, which contains sulfur in its native state, can be produced with selenium substitution. If this amino acid (or its appropriate precursor) is fed to the organisms that produce the protein, pure seleno-methionine-containing proteins can be produced. In usual practice, such proteins are used to solve the crystal structure either as part of multiple isomorphous replacement or, even more directly, by multiple anomalous dispersion experiments on the Se-containing crystals. As the S- and the Se-containing crystals are usually very similar (isomorphous), it is reasonable

to assume that mixed crystals can be produced at any proportion and that the Se-containing proteins will occupy purely random sites. (This is a pure lattice gas; the probability of any one of the perfectly ordered sites in the crystal being occupied by a S-containing or a Se containing protein is the same and no correlation exists between the occupancies of different sites.) Suppose the fraction of Se-methionine is p and that of S-methionine is $1-p$. Such a crystal produces a Bragg diffraction pattern characteristic of the weighted average of the two protein species. In addition, it produces a diffuse diffraction pattern that would be produced by a single molecule that consists of the difference between the two protein species. The integrated intensity of the diffuse pattern is proportional to $p(1-p)N$, where N is the number of molecules in the crystal. As the only difference between the two species of proteins is in the substitution of Se for the S atoms, we would get the diffraction pattern of those difference atoms. The calculation is similar to that described by Jagodzinski and Frey.[49] (In the measurement of the diffuse diffraction pattern, the Bragg peaks must be avoided.)

This would be an unusual hologram. Deciding what is the reference and what is the object is difficult. It could be sampled as finely as needed to satisfy the Nyquist criterion, and the methods outlined in this report could probably be used for the recovery of the atoms. In the language of crystallography, the entire volume of the Ewald sphere is accessible within the $1/\lambda$ limit, not only the reciprocal lattice points. Because it is not concentrated into Bragg peaks, the diffuse intensity is fairly low. But, the contrast of the diffraction pattern is of the order of unity, as opposed to the x-ray holograms described above[11,12] that have a very low contrast. It is also different from holographic LEED[42] based on multiple scattering where the reference wave is the one scattered from a disordered adatom and the object wave is the wave scattered from the ordered part of the crystal after already being scattered at least once by the adatoms. In the experiment outlined above, only singly scattered photons participate. When several molecules are in the unit cell, related by crystallographic (or noncrystallographic) symmetry, the measured diffuse diffraction pattern is the sum of those of the individual molecules. This may make the solution of the molecular structure difficult.

Discussion and Conclusions

Let us remind ourselves that both electromagnetic and matter waves have amplitude and phase information associated with them. When these waves are detected, the phase information is lost in most circumstances. There are two fundamental reasons for this loss. When the number of photons (particles) per mode is less than unity, impossibility of detection of the absolute phase follows from Heisenberg's uncertainty relations. (Note that this is always true for Fermion waves like electrons.) If the wave consists of a very large number of coherent photons per mode, as in a radio frequency (microwave) field or in a good laser, the phase is measurable in principle. Even so, we still need a detector with a short response time compared to the period of the incident wave for the measurement of the absolute phase.

A way around this limitation has been known ever since Young's two-slit experiment. When two-coherent waves interfere, their relative phases influence the detected intensity. The generic name of interferometry is associated with this method. The requirements of a large number of coherent photons per mode and of fast detectors are relaxed, leaving only the requirement of coherence within the phase volume of the interferometer. As an extreme example, a photoelectron can be coherent only with itself. Photoelectron coherence length is determined by its energy spread and by the size of its source; thus the length can be very different in different directions. The volume enclosed by coherence length in all directions is called its coherence volume. In particular, if a photoelectron is emitted from a very narrow inner shell level of an impurity atom by a monochromatic photon, its coherence volume can be fairly large on an atomic scale. If such a photoelectron is scattered from two objects whose vectorial distance is within the electron's coherence volume, the scattered intensity shows an interference pattern.

Gábor realized that if one of these scatterers is a point object, a simple two-stage procedure reconstructs the second scatterer. This was the start of holography. Gábor's original goal was to improve electron microscopy, which was severely limited in its resolution by lens aberrations. In the approximately 50 intervening years, his goal has (almost) been met, mostly by gradual developments in holographic methods of electron microscopy. Our starting point in this report was an observation, made by this author some time ago, that by bringing the reference and the unknown objects close together (to within a few wavelengths), some of the well-known limitations of holography can be ameliorated. Specifically, the coherence length of the waves can be quite small, the recording medium can be of low resolution, and some of the extreme stability requirements are eliminated.

Most of the methods described in this report utilize many identical objects that are oriented the same way. When the recording screen is far compared to the distance of the reference to the object, the resulting diffraction pattern is the same as that of a single object, except that in crystals (for which the Bragg conditions have to be satisfied) it becomes sampled. If successful, the reconstruction produces the image of the single (repeated) object. The obvious advantage of these methods is that they ameliorate the radiation damage to a sample. For many biologically important systems, this is the only way to avoid total destruction of the sample before an image is obtained.

In summary, this report discussed the generalization of Gábor's holographic method to a complicated reference wave. This forced us to reconsider his reconstruction method, to discuss its mathematical properties, to deal with the dual image and to invent new reconstruction techniques. We found a close connection between ordinary interferometry and holography. We also established a similar connection to the general field of inverse problems. We hope the ideas presented in this report will find exciting applications in the future. ▲

Appendix

We summarize some of the relevant discussion on frames, following Daubechies[33] and Heil and Walnut[34] (referred to below as D and H&W, respectively.) To represent the general scatterer density in Eq. 8, $\rho(\mathbf{x})$ and its scattering power, $\tau(\hat{\mathbf{R}}, \hat{\mathbf{k}}_{\text{in}}, \mathbf{x})$, in terms of basis functions, $\Omega(\mathbf{x} - \mathbf{x}_n)$, or their scattering power, $\tau(\mathbf{R}, \mathbf{k}_{\text{in}}, \mathbf{x} - \mathbf{x}_n)$, the latter have to be frames. This is simplier in the x-ray region. The scattering power of materials is proportional to the electron densities of molecules, $\rho(\mathbf{r})$, which are positive functions of limited resolution or bandwidth. According to Eq. 26 they are to be represented by a set of Gaussian basis functions.

A set of functions $\{x_n\}$ is called a frame (with respect to the functions x) if for any one of the functions x, the sum

of the squares of its scalar products with all of the functions x_n is bounded both from above and from below (D 3.2.1, H&W 2.1.1):

$$A\langle x,x\rangle \le \sum_n \left(\left|\langle x,x_n\rangle\right|^2 \right) \le B\langle x,x\rangle; \quad 0 < A,B < \infty, \tag{A1}$$

where the scalar product <x,y> (also called a convolution or projection) is defined as the integral,

$$\langle x,y\rangle = \int x(\mathbf{r})y(\mathbf{r})d\mathbf{r}. \tag{A2}$$

The existence of a frame ensures that the operator S that produces $<x,x_n>$ from x is a bounded linear operator with a bounded inverse operator S^{-1}. The functions x can then be expanded with the help of the operators S and S^{-1}. Indeed, let us define the operator S by

$$Sx = \sum_n \left(\langle x,x_n\rangle \right) x_n . \tag{A3}$$

It follows that the function x can be expanded (represented) in terms of the set of functions x_n by the formula (D 3.2.8, H&W 2.1.5):

$$x = \sum a_n x_n , \text{ where } a_n = \left\langle x, S^{-1}x_n \right\rangle. \tag{A4}$$

The set of functions $\{S^{-1}x_n\}$ is also a frame, called the dual frame. It satisfies the bounds (D 3.2.6, H&W 2.1.4)

$$B^{-1}\langle x,x\rangle \le \sum_n \left(\left|\left\langle x,S^{-1}x_n\right\rangle\right|^2 \right) \le A^{-1}\langle x,x\rangle. \tag{A5}$$

The significance of Eqs. A1 through A5 for us is that if we can show that our basis set of Gaussians Eq. 26 is indeed a frame, we can be assured that any electron density can be represented by them. Moreover, we are assured that the representation is mathematically stable in both directions in the following sense. Given two sets of coefficients, a_n and b_n that are close, the set of coefficients $c_n = a_n - b_n$, when used in Eq. A4, defines a function that is small because in Eq. A1, B < ∞. The converse is that if the function x is small, the set of coefficients is also small because of the right side inequality in Eq. A5. In fact, Eq. A4 supplies an algorithm for the representation of a known electron density.

The most familiar basis function sets are orthonormal. Those are the generalizations of orthonormal (Cartesian). coordinate systems to (infinite dimensional) function spaces. Such basis sets always constitute frames, with $A = B = 1$ in Eq. A1, so frames can be thought of as generalizations of orthonormal basis sets. Frames are usually not orthogonal and are usually redundant in that the representation presented in algorithm A4 is not unique. However, a connection exists among all possible representations of x in terms of x_n (H&W 2.1.5). If, in addition to the representation A4 we can find another set of coefficients that represents x (e.g., by using the EDEN algorithm)

$$\mathrm{x} = \sum c_n x_n, \tag{A6}$$

the following connection exists between the two representations:

$$\sum |c_n|^2 = \sum |a_n|^2 + \sum |a_n - c_n|^2. \tag{A7}$$

Equation A7 shows that algorithm A4 always yields a representation with minimal norm.

Two main classes of frames are discussed in Refs. 33 and 34. The frames in the first class are called Weyl–Heisenberg–Gábor frames; they can also be viewed as windowed Fourier transforms. The second class are wavelet frames. The basis function representation in this report (Eq. 26) belongs to the former class. The relevant formulas will be shown in one dimension, but they apply to three-dimensional lattices as well.

Given a standard Gaussian

$$g(x) = \pi^{-1/4}\exp(-x^2/2), \tag{A8}$$

we can define a set of functions,

$$g_{m,n}(x) = \exp\,(im\omega_o x)\, g(x - nt_O),\ m,n = \text{integer}. \tag{A9}$$

They are the basis functions of the windowed Fourier transform. They measure the frequency content, around $m\omega_o$ in frequency, of a small section of a function centered around nt_O in space. It is shown in D 3.4.4 that if $\omega_o t_o < 2\pi$, the set of functions A9 constitute a frame. The frame bounds A, B of A1 are estimated in D Table 3.3. The significance of the frame bounds is that if they are close, the representation of most functions converges rapidly. In our application we are interested in a representation where only $m = 0$ is kept (Eq. A9).

Let us identify the variables in Eqs. A8 and A9 with those of Eq. 26 as

$$x = 2^{1/2}\,\pi r/d_{\mathrm{res}};\ t_o = 2^{1/2}\,\pi\Delta r/d_{\mathrm{res}}. \tag{A10}$$

The experimental data set and the structure factors of the known part are anodized according to Eq. 29 to have the same inherent resolution as the basis set Eq. 26. It follows that all possible (positive) electron densities have a maximum resolution d_{res} and their Fourier transform falls off in reciprocal space as $\exp(-d_{\mathrm{res}}{}^2\,|\mathcal{F}^T\mathbf{h}|^2)$. A prototypical function with these properties is one of the Gaussians in Eq. A9 corresponding to one of the Gaussian basis functions of Eq. 26. The Fourier amplitudes of such a function can be calculated by the formula

$$\int_{-\infty}^{\infty} \exp\left[-\left(x - nt_O\right)^2/2\right]\exp\left(im\omega_O x\right)dx = (2\pi)^{1/2}\exp\left(-m^2\omega_O{}^2/2\right)\exp\left(imn\omega_O t_O\right). \tag{A11}$$

We will choose

$$\Delta r = d_{\mathrm{res}}/\pi,\ \ \omega_o t_O = \pi. \tag{A12}$$

The frame bounds from D Table 3.3 are close to $A = 1.5$ and $B = 2.5$ with their ratio being about 1.7. Thus the frame is fairly tight, and we should expect fairly good convergence of the representation for any electron density. Moreover, from Eq. A10 we can calculate the value of the first nonzero Fourier component, $\exp(-\omega_O{}^2/2) = 0.085$. Therefore if we neglect all higher Fourier components of the frame, i.e., if we restrict our representation to Eq. 26, the maximum relative error we make is 8.5%. Similarly, the restriction that all the amplitudes of the Gaussians be nonnegative to satisfy the constraint that the electron density be nonnegative everywhere is expected to cause a

similarly small error. This is the mathematical basis of our representation of electron densities.

The above discussion can be generalized to a multiresolution representation either in terms of Gábor frames or in terms of wavelets. In either case satisfying positivity everywhere can be quite difficult.

Acknowledgment. This work was performed in part under the auspices of the U.S. Department of Energy under Contract No. W-8405-ENG-48.

* This work was performed in part under the auspices of the US. Department of Energy under Contract No. W-7405-ENG-48.

† If the reference beam does not pass through the specimen, the number of particles per voxel that have to be detected is reduced by the square root of the ratio of the reference intensity to the scattered intensity. This reduction can be called an interferometric advantage; it is analogous to the well known heterodyne advantage of radio frequency detection.

‡ A function vanishes everywhere outside the domain of its support.

§ The enantiomorph of an object is its three dimensional reflection around a center of symmetry.

** There is a lively debate on super-resolution in the astronomy literature.[26] By super-resolution one usually means resolution finer than λ/NA in the transverse direction and λ/NA^2 in the longitudinal direction, as mentioned in the **Introduction**. The author's position is that super-resolution is possible only if additional information is available on the shapes of the objects to be recovered. Most simply: if we know the internal structure of an object, e.g., that it is point-like, we can find its position to a very high accuracy even from relatively low resolution data—given the data have very good signal-to-noise ratio and the possible systematic errors are also well known. If such knowledge is not available, super-resolution is not possible.

References

1. D. Gábor, A new microscopic principle, *Nature* **161**, 777 (1948).
2. D. Gábor, Microscopy by reconstructed wave-fronts, *Proc. R. Soc.* (London) **A197**, 454 (1949).
3. A. Tonomura, Recent developments in electron holography for phase microscopy, *J. Electron Microsc.* **44,** 425 (1995), and references therein.
4. A. Orchowski, W. D. Rau, and H. Lichte, Electron holography surmounts resolution limit of electron microscopy, *Phys. Rev. Lett.* **74,** 399 (1995), and references therein.
5. S. Lindaas, M. Howells, C. Jacobsen, and A. Kalinovsky, X-ray holographic microscopy via photoresist recording and atomic-force microscope readout, *J. Opt. Soc. Am. A* **13,** 1788 (1996).
6. P. Morin, M. Pitaval, and E. Vicario, Low energy off-axis holography in electron microscopy, *Phys. Rev. Lett.* **76,** 3979 (1996), and references therein.
7. E. N. Leith and J. Upatnieks, Reconstructed wavefronts and communication Theory, *J. Opt. Soc. Am.* **52,** 1123 (1962).
8. D. Sayre and H. N. Chapman, X-ray microscopy, *Acta Cryst. A* **51,** 237 (1995).
9. R. Henderson, The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological specimens, *Quarterly Rev. Biophy.* **28,** 171 (1995).
10. A. Szöke, X-ray and electron holography using a local reference beam, in *Short Wavelength Coherent Radiation: Generation and Applications,* D. T. Attwood and J. Bokor, Ed., American Institute of Physics Conference Proceedings No. 147, New York, 1986.
11. M. Tegze, and Gy. Faigel, X-ray holography with atomic resolution, *Nature* **380**, 49 (1996).
12. T. Gog, M. Len, G. Materlik, D. Bahr, C. S. Fadley, and C. Sanchez-Hanke, Multiple-energy x-ray holography: atomic images of hematite (Fe_2O_3), *Phys. Rev. Lett.* **76,** 3132 (1996).
13. D. K. Saldin, Holographic crystallography for surface studies: a review of the basic principles, *Surface Rev. Lett. Proc. WWEDIS* (1996).
14. C. S. Fadley, Diffraction and holography with photoelectrons and Auger electrons: some new directions, *Surf. Sci. Repts.* **19**, 231 (1993).
15. H. J. Kreuzer, K. Nakamura, A. Wierzbicki, H. W. Fink and H. Schmid, Theory of the point source electron microscope, *Ultramicrosc.* **45**, 381 (1992).
16. L. Greengard, Fast algorithms for classical physics, *Science* **265,** 909 (1994).
17. A. Szöke, Holographic methods in x-ray crystallography, II. Detailed theory and connection to other methods of crystallography, *Acta Cryst.* **A49**, 853 (1993).
18. M. Bertero, Linear inverse and ill posed problems, *Advances in Electronics and Electron Physics,* **75**, Academic Press, 1989.
19. D. G. Luenberger, *Linear and Nonlinear Programming,* 2nd ed., Addison Wesley, Reading, MA, 1984.
20. P. C. Sabatier, in *Basic Methods of Tomography and Inverse Problems,* P. C. Sabatier, Ed., Adam Hilger, 1987.
21. F. Natterer, *The Mathematics of Computerized Tomography* John Wiley & Sons, Chichester, 1986.
22. G. H. Golub, and C. F. Van Loan, *Matrix Computations,* 2nd Ed., Johns Hopkins University Press, Baltimore, 1989, p. 71.
23. C. Lánczos, *Linear Differential Operators,* Van Nostrand, London 1961.
24. R. W. James, *The Optical Principles of the Diffraction of X-Rays,* Reprinted by Ox Bow Press, Woodbridge, 1982.
25. G. Bricogne, in *Int. Tables for Crystallography,* **B**, U. Shmueli, Ed., Kluwer Academic Publishers, Dordrecht, 1992.
26. J. Nunez, Ed., Special issue: image reconstruction and restoration in astronomy, *Intern. J. of Imaging Syst. and Technology,* **6,** 195 (1995).
27. J. J. Barton, Photoelectron holography, *Phys. Rev. Lett.,* **61,** 1356 (1988).
28. S. Y. Tong, Hua Li, and H. Huang, Energy extension in three-dimensional imaging by electron emission holography, *Phys. Rev. Lett.* **67,** 3102 (1991).
29. J. J. Barton, Removing multiple scattering and twin images from holographic images, *Phys. Rev. Lett.* **67,** 3106 (1991).
30. A. Szöke, Electron-diffraction spectroscopy and the holographic inverse problem, *Phys. Rev.* **B47,** 14044 (1993).
31. D. K. Saldin, X. Chen, N. C. Kothari, and M. H. Patel, Atomic position recovery by iterative optimization of reconstructed intensities: overcoming limitations of holographic crystallography, *Phys. Rev. Lett.* **70,** 1112 (1993).
32. J. C. Dainty, and J. R. Fienup, *Image Recovery: Theory and Applications,* H. Stark, Ed., Academic Press, orlando, 1987, Chap. 7.
33. I. Daubechies, *Ten Lectures on Wavelets,* SIAM, Philadelphia, 1992.
34. C. E. Heil and D. F. Walnut, Continuous and discrete wavelet trans forms, *SIAM Review* **31,** 628 (1989).
35. G. J. Maalouf, J. C. Hoch, A. S. Stern, H. Szöke, and A. Szöke, Holographic methods in x-ray crystallography, III: First numerical results, *Acta Cryst.* **A49,** 866 (1993).
36. P. Béran, and A. Szöke, Simulated annealing for phasing using spatial constraints, *Acta Cryst.* **A51,** 20 (1995).
37. J. R. Somoza, H. Szöke, D. M. Goodman, P. Beran, D. Truckses, S.-H. Kim, and A. Szöke, Holographic methods in x-ray crystallography, IV: A fast algorithm and its application to macromolecular crystallography, *Acta Cryst.* **A51,** 691 (1995).
38. A. Szöke, H. Szöke, and J. R. Somoza, Holographic methods in x-ray crystallography, V: Multiple isomorphous replacement, multiple anomalous dispersion and non-crystallographic symmetry, *Acta Cryst.* **A53,** 291 (1997).
39. L. A. Amos, R. Henderson, and P. N. T. Unwin, Three-dimensional structure determination by electron microscopy of two-dimensional crystals, *Prog. Biophys. Mol. Biol.* **39,** 183 (1982).
40. W. Kühlbrandt, D. N. Wang, and Y. Fujiyoshi, Atomic model of plant light-harvesting complex by electron crystallography, *Nature* **367,** 614 (1994).
41. C. J. Gilmore, K. Shankland, and G. Bricogne, Applications of the maximum entropy method to powder diffraction and electron crystallography, *Proc. Roy. Soc. London, A* **442,** 97 (1993)
42. D. K. Saldin and P. L. de Andres, Holographic LEED, *Phys. Rev. Lett.* **64,** 1270 (1990).
43. J. Spence, W. Qian, and X. Zhang, Contrast and radiation damage in pointprojection electron imaging of purple membrane at 100 V, *Ultramicros.* **55,** 19 (1994).
44. J. M. Rodenburg and R. H. T. Bates, The theory of super-resolution electron microscopy via Wigner-distribution convolution, *Phil. Trans. R. Soc. Lond. A* **339,** 521 (1992).
45. P. D. Nellist, B. C. McCallum, and J. M. Rodenburg, Resolution beyond the information limit in transmission electron microscopy, *Nature* **374,** 630 (1995).
46. H. N. Chapman, Phase-retrieval x-ray microscopy by Wigner-distribution deconvolution, *Ultramicrosc.* in press (1997).
47. J. Kirz, C. Jacobsen, and M. Howells, Soft x-ray microscopes and their biological applications, *Quart. Rev. Biophys.* **28,** 33 (1995).
48. G. Xu, Atomic resolution hologram, *Appl. Phys. Lett.* **68,** 1901 (1996).
49. H. Jagodzinski and F. Frey, Disorder diffuse scattering of x-rays and neutrons, in *Int. Tables for Crystallography.* Vol. *B,* U. Shmueli. Ed., Kluwer Academic Publishers, Dordrecht, 1992.