Optical Disks for Image Storage and Distribution

Douglas G. Stinson*

Digital and Applied Imaging, Eastman Kodak Company, 460 Buffalo Road, Rochester, New York 14652-3836

The fundamental properties of light result in images being an important part of peoples' lives. One of the reasons for their importance is that images are information rich; in the digital world this means large file sizes. These same properties of light allow optical disks to store information at high density. The characteristics of optical disk systems are well matched to those of digital images, particularly when high-quality digital images needed to be shared between, or distributed to, a number of individuals. Advances in many areas, from materials to electronics to coding theory, as well as high-volume production, has provided this capability to the average personal computer user.

Journal of Imaging Science and Technology 42: 39–48 (1998)

Introduction

Optical disks,[†] for example, CD-ROMs, are widely used to store and distribute images. Images on an optical disk are generally represented in digital form and, as Negroponte, founding director of the MIT Media Lab., has said, "bits are bits."[1] The content of an optical disk corresponds to an image only by convention; lose the "Rosetta Stone" that defines that convention and the content could just as well be a bank account, or even noise.[‡,2] Consequently, a discussion of the issues relevant to optical disk technology can take place without reference to imaging. Conversely, the relevant technical issues in digital imaging can be discussed without reference to whether an image exists on an optical disk, on a magnetic floppy disk, or as a string of electromagnetic pulses propagating through the ether. This independence is one of the strengths of digital modalities.

Nonetheless, certain characteristics of optical disks make them the ideal medium for the storage and distribution of images. Optical disks excel at storing and retrieving large, contiguous blocks of information inexpensively using systems whose tolerances are relatively loose (allowing information to be recorded on one device and reliably played back on another). Images are large, contiguous blocks of information. The properties of both optical disk systems and image files are derived from the fundamental properties of light.

The energy levels of atoms and molecules are of the same scale as the energy of the electromagnetic radiation we call light. So after light interacts with matter, it contains a wealth of information of great potential relevance to us creatures who live or die based on how we interact with aggregations of those atoms and molecules. The spatial resolution of that information is on the order of the wavelength of light, a few hundred nanometers. At that resolution, information can be carried as intensity, wavelength, polarization, phase, and the time variation of those properties. Because light propagates through space, that information is available at a distance (of great value if that information is "predator nearby"!). Because of the similarity in energy scales of molecules and light, it is fairly easy to construct objects that manipulate light (e.g., a lens), or detect it (e.g., the retina). Because of their usefulness and ease of construction, image-forming organs have evolved independently perhaps 100 times in life's history.[§,3] While only 6 out of 30 phyla have image-forming eyes, such eyes confer so many evolutionary advantages that these phyla dominate, comprising 95% of all animals.[4] We have come to prize information-rich images to the point where we will, by preference, represent the most abstract concept as an image.

These same properties of light are exploited for optical recording. Because the energy scale is appropriate, light can be used to modify a material in a way that represents information. Similarly, light will interact with and, therefore, be modified by the information storage material. Features can be created and detected in the storage material with spatial resolution on the order of the wavelength of light. This enables a great deal of information to be stored in a small amount of space. Because light propagates through space and can be manipulated by fairly simple means, the storage material can be at a distance from the rest of the storage system. This makes it fairly easy to build a system where the storage medium can be moved from device to device, while still reliably detecting the submicron features that represent the stored information.

[†] At least in the United States, the preferred spelling of the word meaning any thin, flat circular plate or object is "disk." However, the official name for the CD is Compact *Disc*. In this paper we will use the convention, wherever possible, of using "disk" to refer to a generic optical disk and "disc" to refer to members of the compact disc or DVD family of products.

[‡] A basic tenet of information theory is that, when encoded for maximum efficiency (i.e., of greatest information transferred per unit time, or stored per unit of space), the information appears random, in the sense that any symbol (e.g., 1 or 0 for a given bit) is equally probable and can not be predicted from the value of the preceding symbols.

[§] Recent discoveries have re-energized the controversy over what has become the conventional interpretation of the evolution of the eye. (See, Ref. 4)

Original manuscript received July 21, 1997

[*] Current address: Seagate Recording Media, 47050 Kato Road, Fremont, California 94538

Based on this description, it would seem that an optical disk would be an ideal storage medium for any type of information. While many applications exist for optical storage, the characteristics of images and optical disks make them well suited to each other. Two general tasks need to be performed by any data retrieval device: locating the data of interest and then reading that data. Most large datasets are really aggregations of smaller "chunks" of data. A massive bank record will be composed of many small datasets corresponding to individual customers, for example. In this situation, much of the activity of a data retrieval system will consist of skipping from one small dataset to another. Therefore a system suited to this type of information needs to minimize the time spent on the "access" task by rapidly accelerating the transducer used to read the data and moving the transducer quickly to the location of the dataset where ever on the storage medium it may be. In an optical data storage system, the transducer includes the object used to manipulate the light—the lens—which is a large block of glass. It takes time to get such a massive object up to speed. Consequently, for the system to be efficient, one wants to minimize the number of "access" tasks. If the dataset is large, this will be the case. Images are typically large, contiguous blocks of data and so fit this criteria nicely.

Types of Optical Disks

Optical disks fall into three broad categories: Read Only Memories (ROM), Write Once—Read Many times (WORM), and rewritable.

In ROM disks the information is built into the disk at the time of manufacture. The disks are distributed to be read using an optical disk reader by the end user. ROM disks are suitable when copies of identical information need to be distributed to many users. Examples include the ubiquitous CD Audio and CD-ROM discs.

WORM disks are manufactured with a light-sensitive recording layer. The information is recorded on to the disk by the user using an optical disk drive. Once recorded, the information cannot be altered without damaging the disk. In essence, the user, rather than the manufacturer, creates a ROM disk. WORM disks are useful for distributing one or a few copies of large datasets, where it is important that any alteration of the data be detectable (e.g., if there needs to be an audit trail of changes), or for archiving data for long-term storage. Because WORM disks are often less expensive that rewritable optical disks, they are also used for making back-up copies of information stored on a computer's hard drive. Even when continued "on line" access to the data is required, if the data changes infrequently it is often less expensive to store it on a WORM diss than on a rewritable disk. The CD-R (Compact Disc—Recordable) is the most common WORM disk.

Rewritable optical disks are more similar to the magnetic hard disk or floppy disks familiar to all personal computer users than are the optical disks just discussed. Information on a rewritable optical disk can be replaced with new information by the user. The most common rewritable optical disks are the 5.25" and 3.5" magneto-optic disk systems. These allow essentially unlimited read–write–erase cycles. Until recently these devices suffered from long access times because two rotations of the disk were required to write data—one to erase the old data and one to write the new. Systems based on erasable phase change technology, such as Matsushita's PD or the new CD-RW (Compact Disc—Rewritable) eliminated this delay by using "direct overwrite," but at the cost of only allowing a few thousand write–erase cycles. Magneto-optical systems now feature direct overwrite as well.[5,6]

Disks for Images

Although every form of optical disk is used to store or distribute images, three are dominant: video, compact (particularly in the form of Photo CD and FlashPix CD), and the recently available digital video discs (DVD).

Video Disk. Video disk is a system for storing and distributing video motion images, such as movies, for display on conventional television sets. It was first demonstrated by Philips and MCA in 1972 and has been on the market since 1978. Video disks are a type of ROM disk, but are unique among popular optical storage systems in that the information is stored in analog, as opposed to digital, form. They exploit the characteristics of the television video signal that consists of a carrier frequency modulated according to the image information: the signal is a sine wave whose period varies with time.

The information representing the video signal is stored as depressions ("pits") in the 12- or 8-in. diam. polymethylmethacrylate (PMMA) substrate comprising the disk. The side of the disk with the pits is metalized with aluminum or an aluminum alloy to make it reflective, and two such disks are laminated together, with the reflective layers forming the center of the sandwich.

The edges of these pits are located at the zero crossings of the frequency modulated sine wave. The pits vary in length continuously and represent the instantaneous period of the video signal. As the disk rotates under a focused laser spot, these pits modulate the intensity of the reflected light by a mechanism to be described later. Because the pits are on the order of the size of the laser spot, even though the pits are roughly rectangular in cross section, the reflected light intensity varies sinusoidally with time. Converting light intensity to electric current recovers the original video signal.

Because the pit lengths vary continuously, the information density on a video disk is very high. But because the lengths are not discrete as they would be in a digital system, no discrimination between signal and noise is possible; noise is as faithfully reproduced as the video signal. Because all frequencies (within the video bandwidth) are possible in the original signal and this signal is simply transcribed to the disk, all possible combinations of pit lengths are possible as well. No way exist to discriminate between a defect on the disk and a legitimate video signal. For video playback, this is generally acceptable because typical defects are small, affecting a very small region of the image and are fleeting in time. Persistence of vision, which allows us to perceive discrete image frames as continuous motion, also allows us to ignore all but the most egregious of these defects.

Compact Discs. A different type of ROM disk, compact disc—digital audio (CD-DA), was introduced by Sony and Philips in the fall of 1982 as a method of conveniently and inexpensively distributing high-fidelity audio. Because the information was stored in digital form and "bits are bits," the format was quickly expanded to include digital data (CD-ROM) in 1984, video (CD-V), interactive games and multimedia (CD-I) in 1987, and still images (Photo CD) in 1990. With an estimated 100 million CD-ROM readers and over half a billion CD audio players, the Compact Disc is easily the most successful optical storage system and, in fact, one of the most successful consumer electronic products of all time.

On a CD, as on a video disk, the information is stored as "pits" in a substrate. The 120-mm-diameter, 1.2-mm-thick CD substrate is made of polycarbonate, and the side of the substrate with the pits is coated with aluminum or an aluminum alloy. Unlike the video disk, a CD is single sided, with the reflective layer protected by a coating, generally a UV cured acrylate. A schematic cross section of a compact disc is shown in Fig. 1.

Approximately 650-Mbyte of data, or 74 min of sound, can be stored on a CD.

Photo CD. Jointly developed by Kodak and Philips, Photo CD[7,8] was originally positioned as a method for consumers to store their "snapshots" conveniently and share them with family and friends by viewing the pictures on the family television. It found its greatest use, however, in the graphic arts industry where it has become a standard method for inexpensively converting photographs to digital form at high resolution and with excellent color fidelity. Once on a Photo CD, these images are easily exchanged between the various parties involved in the production of, for example, an advertisement in a magazine.

Because few copies of the disc are required, the mass production techniques used for CD-Audio and CD-ROM are not appropriate. Instead, Photo CD became one of the first applications for recordable CD (CD-R) technology. CD-R, in the form in which it exists today, was first developed by Taiyo Yuden.[9–11] The information molded into a conventional CD is replaced by a continuous spiral groove. This defines the location at which data will subsequently be recorded. A recording layer comprised of an organic dye is interposed between the substrate and the reflector layer. To achieve sufficient reflectivity at the operative wavelength of 780 nm and to ensure a long-lived disc, gold replaces aluminum as the reflector layer. For Photo CD, the cyanine dye mixtures used for the recording layer in many CD-R discs are replaced by a phthalocyanine dye to improve light stability and longevity.

Images are stored on a Photo CD by first scanning a 35-mm negative at a resolution of 3072 × 2048 pixels. This corresponds to 85 pixel/mm. For comparison a 600-dpi laser printer uses approximately 24 pixel/mm. Eight bits are used to digitize the red, green, and blue intensities (24-bit color). This means each frame of a 35-mm negative produces more than 18 Mbyte of information. These parameters were chosen because studies showed that an 11 × 14" print recreated from this data when viewed from a distance of 13 to 20" could have perceived quality comparable to an optical print of the negative.

The color information is converted to a color model called PhotoYCC. The transformation between red, blue, and green and PhotoYCC is nonlinear to make more efficient use of the bits by using the known response of the human eye as well as the response of various hard copy and display devices. It also encodes color as luminance (Y, a weighted average of red, blue, and green) and two chroma values (Cu = blue – Y and Cv = red – Y). Because the human eye perceives light intensity variation at higher resolution than color, this allows the color information to be stored at lower resolution with virtually no loss in perceived image quality.

Because 18-Mbyte files can tax computer systems and the full-resolution image is not always required, the images are stored on the disc in files called Image Pacs, containing Base/16, Base/4, Base, 4 Base, and 16 Base versions of the images. The 16 Base version corresponds to full resolution, the so-called "digital negative" (see Fig. 2). As suggested by the name, each of the other components of the image pac is formed by averaging adjacent pixels, reducing the resolution by a factor of 2 in both the horizontal and vertical directions. The Base version is used for television display.

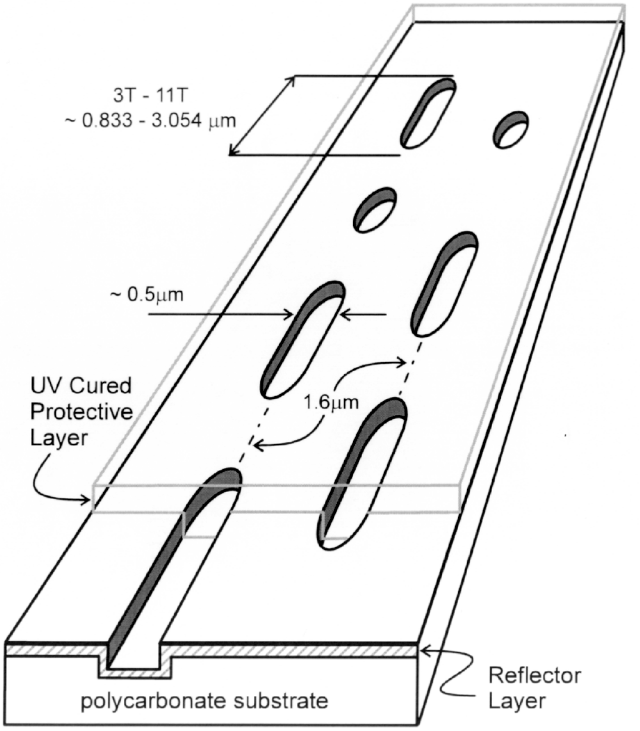

Figure 1. A CD Audio or CD-ROM disc consists of a polycarbonate substrate into which pits are molded, as shown in this cross section. The substrate is coated with a metalic reflector layer and then a UV-cured acrylate protective layer. In a CD-R disc the track of pits is replaced with a continuous groove and an organic dye recording layer is coated on the substrate prior to the coating of a gold reflector layer. A DVD disc would look very similar, except the molded features would be smaller and two substrates would be laminated together.

The three lowest resolution components of the image pac are stored in uncompressed form. This allows for quick retrieval by unsophisticated playback systems. For the higher resolution components, only the difference between that resolution image and the next lower resolution image is stored. These differences are also subjected to a lossless compression scheme known as Huffman encoding. To obtain, for example, the 4 Base image, the Base image is expanded by a factor of 2, the 4 Base component is Huffman decoded and added to the expanded Base image. This process is repeated to obtain the 16 Base image.

Using this scheme, the average size of an image pac is 4.5 Mbyte. However, depending on film type and scene content, the actual size of an image pac can vary between 3 and 6 Mbyte. Typically 100 images can be stored on a Photo CD. Because many photofinishers have the capability of transferring images to Photo CD, this service is now readily available at a cost similar to that of having photographic prints made. Prior to Photo CD, digital images of this quality could only be obtained at special service bureaus and at a cost in the hundreds of dollars per image.

FlashPix CD. Photo CD continues to demonstrate its usefulness for the soft display of images and for high-end applications, for example, in publishing. However, even with the image pac structure, many home computers lack sufficient power to manipulate Photo CD images.

A new digital image format, known as FlashPix, was created to remove this barrier to the widespread use of digital images. The technology incorporated in FlashPix not only stores the image in multiple resolutions, but each resolution is divided into tiles 64 pixel square. Applications can copy from disc to memory only the appropriate portion of

Figure 2. The resolution of the images stored on a Photo CD justify the term “digital negative.” The top image is reproduced from a 5" × 7.5" thermal dye sublimation print of the 4 Base component of the image pac. In the bottom image, that portion of the 16 Base component corresponding to the subject’s eye was enlarged approximately 12× and printed. Close examination reveals the reflection of the photographer and two bystanders.

the image and only the appropriate resolution. No restriction exists on image size (within the system addressing constraint of 2^{32} pixels) or aspect ratio. JPEG (Joint Photographic Experts Group) image compression is optional and multiple color models (in addition to Photo YCC) are supported.

Perhaps the most significant feature of the FlashPix format is the way it treats image manipulations. A variety of image adjustments, including cropping, filtering, spatial orientation, scaling, shearing, color correction, and contrast adjustment are handled by matrix algebra using the concepts of homogeneous coordinate systems and affine transformations.[12] A general spatial operation can be represented by

$$\begin{bmatrix} x' \\ y' \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & 0 & a_{14} \\ a_{21} & a_{22} & 0 & a_{24} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \\ 1 \end{bmatrix}, \tag{1}$$

while color modification can be represented by

$$\begin{bmatrix} Luma' \\ Chroma_1' \\ Chroma_2' \\ 1 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ T_{31} & T_{32} & T_{33} & T_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Luma \\ Chroma_1 \\ Chroma_2 \\ 1 \end{bmatrix}, \qquad (2)$$

For example, the translation of a point (x,y) by an amount dx in the x direction and dy in the y direction can be represented by

$$\begin{bmatrix} x' \\ y' \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & dx \\ 0 & 1 & 0 & dy \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \\ 1 \end{bmatrix}. \qquad (3)$$

To edit an image, a relatively low-resolution version of an image may be loaded into the computer. This takes much less time than loading the full-resolution image and consumes fewer system resources (i.e., memory). Each time the user modifies the image, the elements of the 4 × 4 matrix representing the change are calculated. When the work is complete, the product of the matrices corresponding to each change—known as a viewing parameter—is written to the disc, rather than the modified image itself. If the user wants to print the image, the viewing parameter is applied to the high-resolution version of the image prior to the data being sent to the printer.

Multiple versions of an image can be stored by having the original version, plus a collection of viewing parameters, recorded on the disc. This takes much less space than storing different versions of an image.

Digital Versatile Disc. The Digital Versatile Disc, or DVD, is the latest optical disc system to enter the market. It is unique among digital optical storage systems in having been developed specifically for the distribution and playback of images, in this case, feature movies. The requirements for the system were developed by the DVD Advisory Group, comprised mainly of representatives of movie studios. They recommended that the system be capable of holding a 135-min movie on a single side of a disc, have video quality better than laser disk, 5.1 channel audio with "CD quality," 3 to 5 language soundtracks, at least one subtitle channel, multiple-image aspect ratios, copy protection, and parental lock feature at a cost similar to CD.

Many people recognized the success the CD format achieved by leveraging the format across many applications. An ad-hoc "Technical Working Group" was formed from representatives of computer system companies. This group recommended that DVD also have a single format for computer and TV-related applications, be able to read existing CD discs, the read-only disc systems be compatible with future read-write and write-once discs, a single file system for all content and disc types, and high performance for sequential and random-access data types.

To meet the requests of both the movie and computer industries, three members of the DVD family have been introduced: DVD-Video, DVD-ROM, and DVD-RAM. DVD-Video allows movies to be played back on a television set. DVD-ROM is a computer peripheral, basically a higher capacity version of a CD-ROM that will also play movies. DVD-RAM is a rewritable optical disc for use as a computer peripheral. The goal of complete compatibility—defined as the ability of one format's hardware to read the other format's disc—between the ROM and RAM formats was not fully achieved in the first products to be marketed. However second-generation DVD hardware has been announced that remedies this problem. It is an indication of the speed at which this industry moves that the second generation was announced before the first generation had fully penetrated the sales channels.

A DVD disc consists of a 120-mm-diameter 0.6-mm-thick polycarbonate substrate into which digital information has been molded in the form of pits. The side with the depressions is coated with an aluminum or aluminum alloy reflector layer. A second 0.6-mm-thick substrate is then laminated to the side of the disc coated with the reflector. The second substrate may optionally be molded with additional information to create a double-sided disc.

The length of the pits and the spacing between tracks (0.4 and 0.74 µm, respectively) is reduced compared to CD where the corresponding values are 0.83 and 1.6 µm. More efficient data coding is also employed. This allows a DVD-Video or DVD-ROM disc to hold 4.7 Gbyte of information. DVD-RAM discs currently hold 2.6 Gbyte, the decreased capacity being the price paid for rewritablility.

The image information has a resolution of 720 × 480 for NTSC (720 × 576 for PAL) and is encoded using MPEG-2 compression. (MPEG stands for "Motion Picture Experts Group.") The images can have either a 4:3 or 16:9 aspect ratio.

Basic Principles of Optical Disks

The functions that must be performed in a generic optical disk system are given in the following outline:

1. Modify the incoming signal to match the characteristics of the data storage channel.
    - Error correction coding
    - Append system "book keeping"
    - Modulation coding
    - Pre-emphasis
2. Modify the recording material according to the information signal
    - Create the "optical stylus"
    - Address the appropriate area with the stylus
    - Expose the recording material
    - Form optically detectable features
3. Detect the features
    - Scan the disk with the optical stylus
    - Modify a property of the light according to the recorded features
    - Convert the change in properties of the light into electrical signals
4. Recover the digital data
    - Post-emphasis (equalization)
    - Analog-to-digital conversion
    - Demodulate
    - Error Detection and Correction

Advances in optical recording technology are typically paced by steps 2 and 3, specifically laser and recording materials technology. Steps 1 and 4 use existing analog and digital electronic technology to best exploit whatever characteristics the optics and materials provide. If the electronics limit systems performance it is because of cost, size, power consumption, or other practical considerations, not fundamental technology. Consequently, to understand optical recording systems it is best to work from the "inside out."

The Optical Stylus. The descriptive term optical stylus is used to refer to the focused spot of light used to write and read marks on the surface of an optical disk. The size of this stylus fundamentally determines the storage capacity, while the energy density determines the write data rate of a optical storage device. The light intensity distribution in the stylus is given by the Fourier transform of the intensity distribution in the entry pupil of the focusing lens. The light distribution from a laser is approximately Gaussian and so the distribution in the stylus is as well. The size of the focused spot is given by[13]

$$\sigma \cong 0.25\lambda/\mathrm{NA}. \tag{4}$$

where σ is the diameter of the contour corresponding to the standard deviation of the gaussian light intensity distribution, λ is the wavelength of the light, and *NA* is the numerical aperture of the focusing lens. (The *NA* is the sine of the angle formed by a marginal ray and the optic axis.) Another common measure of the size of the optical stylus is the full width at half maximum intensity FWHM = 2.35σ. In a CD system, λ=780 nm and NA = 0.45, so $\sigma \cong 0.43$ µm. For DVD, $\lambda = 650$ nm and NA = 0.60, so $\sigma = 0.27$ µm. Therefore, one would expect to be able to store $(0.45/0.27)^2 = 2.8$ times more information on a DVD disc than on a CD disc because of decreased size of the optical spot alone.

The size of the spot increases away from the focal point. The distance over which the spot size does not change significantly is called the depth of focus and is given by[14]

$$\Delta z \cong \pm 0.4\lambda/\mathrm{NA}^2. \tag{5}$$

To maintain the performance of the optical disk system the lens-to-disk distance must be maintained to within the depth of focus. For a CD system, this is approximately ±1.5 µm, for DVD ±0.7 µm. Optomechanical systems[15,16] exist to provide this autofocus function, but as higher data storage densities are sought through the use of a shorter wavelength or, more critically, through the use of a higher NA lens, the demands on this system become ever more severe. Nonetheless, this is a significant improvement over magnetic recording technology that must maintain the magnetic head within microinches (≈0.03 µm) to resolve features of comparable in-track length.

Marking the Media. During the recording process, digital information is used to switch the laser between low and high power. In current optical recording systems, the power delivered to the disk is on the order of 1 mW (low-power "read" state) to ≈10 to 20 mW (high-power "write" state). At least one layer of material on the optical disk is designed to absorb light at the laser wavelength. As a result, that layer is heated in the local area illuminated by the optical stylus. The region heated above a threshold temperature undergoes some change in its optical properties. This change could be the formation of a physical hole in the layer,[17–19] bubble formation, a reversal of the direction of magnetization in a magnetic thin film detectable by the magneto-optic Kerr effect,[20,21] dye fade, or an amorphous to crystalline phase transition.[22] Figure 3 shows an atomic force micrograph of the marks formed in a Photo CD.

The marking process is addressed in detail by several authors.[23] Discounting thermal diffusion, the temperature profile of the illuminated area on the disk in the radial direction follows the light intensity profile. The width of the mark will be determined by the point at which the marking threshold temperature is exceeded, which in turn is determined by the width of the optical stylus and the peak light intensity. In the tangential direction, the point at which the marking threshold temperature is exceeded will be determined by the integrated exposure received, which will be determined not only by the light intensity and stylus size, but by the distance over which the stylus is scanned by the rotation of the disk. Thus, marks of various lengths can be created by controlling the duration for which the laser is pulsed to high power. In most current optical storage systems, information is encoded in the relative position of the edge of the mark, so precise control of mark length is critically important.

For WORM and rewritable disks, where the recording is done by the end user of the disk, this recording layer is an integral part of the disk structure. For ROM disks, the recording is typically done in photoresist at the factory at which the disk is manufactured, a process called mastering.[24,25] The photoresist is developed and coated with metal subsequently removed intact. This metal, called a stamper, has protrusions everywhere the photoresist was exposed by the laser. The stamper is used as one face of a mold that is injected with molten plastic. The plastic solidifies to create the disk, which now contains depressions where the photoresist was exposed.

Detecting the Features. To recover the stored information, the disk is scanned under the optical stylus. As the stylus passes over a recorded feature, some property of the light is changed—amplitude, phase, or polarization—which eventually results in a modulation of light intensity on the detector. If the recording mechanism is the micromachining of a hole in a reflective thin film, the mechanism of light intensity modulation is straight forward: the holes are less reflective than the surrounding metal film. The method of light modulation is the same when the recording mechanism is amorphous to crystalline phase transition, the amorphous form of a material being less reflective than the crystalline form.

In other cases the modulation is accomplished through less direct means. In the case of a CD-ROM or video disk, for example, the recorded information exists in the form of depressions ("pits") in the surface of a transparent plastic disk substrate. The information-carrying surface is metalized to make it reflective, and the optical stylus probes the information-carrying surface through the substrate. Clearly, the reflectivity of the pits and the surrounding region (the "land") is the same. However, the pits are smaller than the diameter of the optical stylus and are approximately λ/4 deep. The light reflected from the bottom of a pit is shifted in phase by λ/2 relative to the light reflected from the land. As a result, light reflected from the two areas destructively interferes, making the pit appear dark.

In many optical recording materials, both reflectivity and phase modulation occurs. This can lead to some interesting and counter intuitive results. For example, when the recording mechanism is dye fade, the recorded features absorb less light than the surrounding areas, so one would expect their reflectivity to be higher. However, the change in the real part of the index of refraction that accompanies dye fade can result in a phase shift similar to that described above for a CD-ROM, and the recorded features appear dark. This is the mechanism for recording and reading in cyanine dye-based CD-R media.

In the case of a magneto-optic material, it is the polarization of the light that is modified. In its simplest implementation, the light is linearly polarized and the recording medium is magnetized in a direction normal to the surface. The magnetization is either "up" or "down" according to the recorded information. The direction of polarization of the light reflected from a surface is rotated clockwise or counterclockwise depending on the direction of magnetization, a phenomenon known as the magneto-optic Kerr effect. Before impinging on a detector, the light

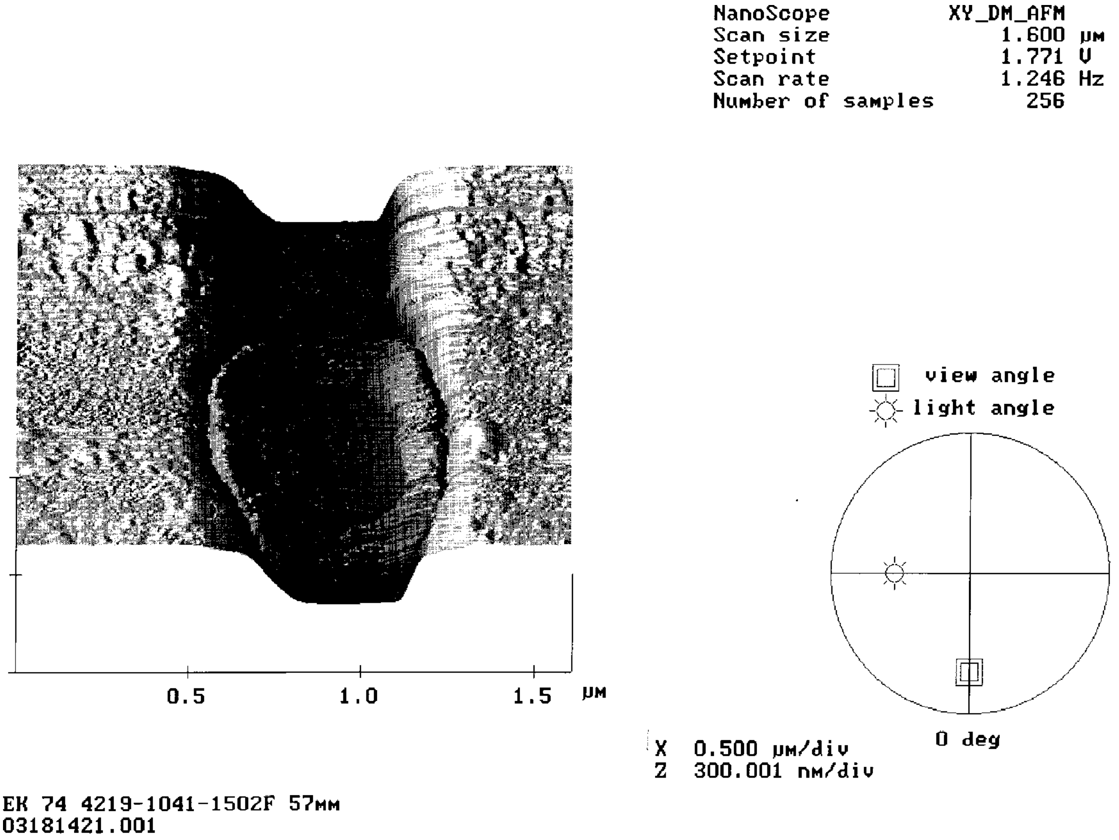

Figure 3. An atomic force microscope image of a recorded mark in a Photo CD disc. The groove is on the order of 150 nm deep. Prior to imaging, the gold reflector layer of the disc was stripped and the dye in the unrecorded regions removed with a solvent wash. This process also caused the collapse of the vesicle of thermally modified dye that constitutes the recorded mark. The robustness of the mark is one of the factors contributing to the light fastness and longevity of data stored in phthalocyanine dye-based CD-R media.

is made to pass through a polarization analyzer, oriented so the direction of rotation causes more or less of the light to be transmitted.

The amount of change in the reflected light—be it amplitude, phase, or polarization—and thus the change in light intensity on the detector is proportional to the amount of change in the optical properties of the marked area on the medium and the area of the mark weighted by the intensity profile of the focused spot—the convolution of the focused spot intensity profile with the marks on the disk. For marks longer than the size of the optical stylus, the signal from the detector will describe a square wave in time. This is because the stylus falls either entirely within a mark or entirely within the intervening land. The only exception is when the stylus passes over the edge of the mark. Here, the apparent edges of the square are rounded by the gaussian shape of the focused spot intensity. When the marks and intervening lands are on the order of the size of the laser spot, even if the marks are roughly rectangular in cross section, the convolution process results in the reflected light intensity varying sinusoidally with time. This is because at any instant, a significant portion of the light will be illuminating both mark and land areas. (In other words, the modulation transfer function of the optical system used to recover the data represents a low-pass filter that strongly suppresses the high-frequency components of the features on the disk.) As the marks and land are made progressively smaller, the amplitude of the sine wave will decrease, decreasing to zero when the marks and lands are always equally illuminated, a situation which is only truly obtained when the size of the marks is zero.

Detecting the marks becomes a problem of detecting what portion of the waveform lies above the inflection point (or zero crossing). The accuracy with which this can be accomplished depends on the signal-to-noise ratio. For sufficiently high signal-to-noise ratio, marks smaller that the size of the optical stylus (as measured by FWHM) can be detected. For example, in a CD-ROM the minimum mark size is 0.83 µm while the stylus FMHM$\cong$ 1µm. For marks of this size the amplitude of the signal can be less than half that of the signal from long (>>1 µm) marks. Noise can result from a variety of sources: spatial variation in the optical properties of the recording medium (media noise), thermally induced fluctuations in the gain of electronic amplifiers (electronic noise), or the statistical nature of light itself (shot noise). In general, it is possible to design an optical disk system so that media noise is dominant.

To better understand the problem of detecting the features on a optical disk, we will consider three cases. In the first are the following sequence of features on the disk: Lm, Sl, and Lm. The capital L or S corresponds to "long" or "short", respectively. The lower case m or l corresponds to "mark" with 0% reflectivity or "land" with 100% reflectivity, respectively. Assume that long features are significantly larger than the optical stylus while the short features have a length approximately equal to that radius of the optical stylus containing half the light intensity. When the stylus is centered on either Lm, no light is returned to the detector. When the stylus is centered on Sl, about half the light falls on the land but half the light falls on the adjacent marks. In this case about 50% of the light is returned. If the noise corresponds to $\leq$ 50%, we

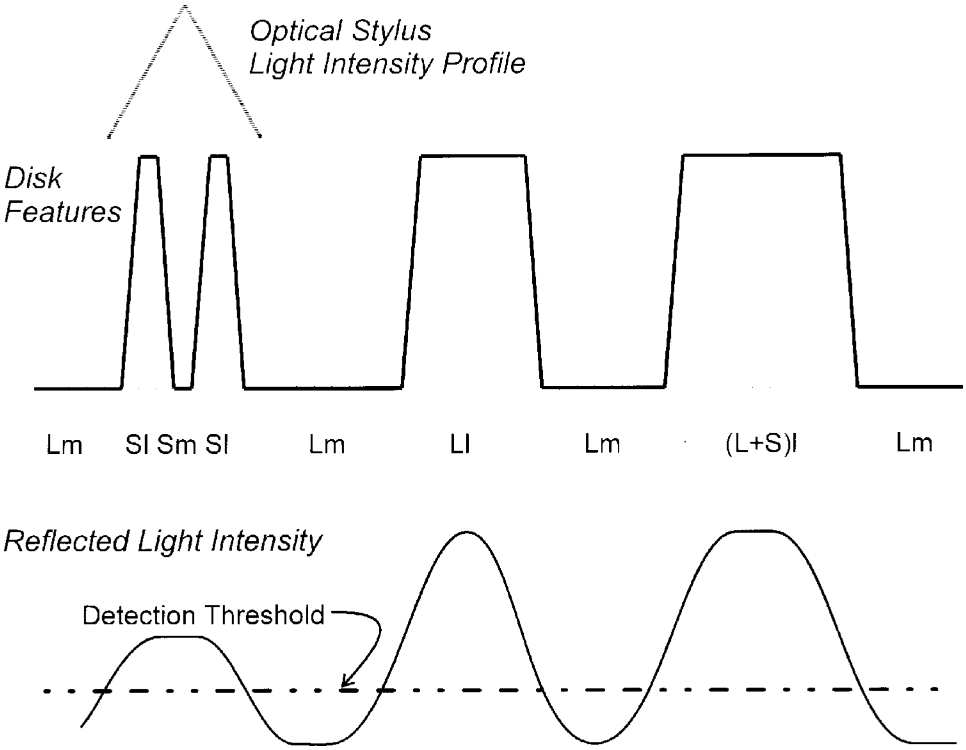

Figure 4. A small change in the length of a feature recorded on an optical disk can be detected, even when a feature of the same size as that change in length cannot. The uppermost portion of the figure shows schematically the light intensity of the optical stylus as a function of position along the track of an optical disk. While the profile of an actual optical stylus is approximately gaussian, a simple triangle function is used to illustrate the phenomenon. In the center of the figure, the reflectivity of the disk as a function of in-track position is plotted. The "L" and "S" correspond to "long" and "short" features, respectively, while "m" and "l" refer to "marks" and "lands". In this example S is 1 and L is 6 units in length. The lower portion of the figure shows the intensity of the reflected light as the optical stylus is scanned along the track. This is calculated by convolving the stylus with the disk reflectivity. The detection threshold is set at 25% of maximum reflected light intensity, which is half the intensity of the light reflected from an isolated short mark. As shown, a short mark between two short lands cannot be detected.

can set a threshold at 25%. Any signal greater than this corresponds to a land, and any signal less than this corresponds to a mark. The Sl feature is easily detected.

The second and third cases are illustrated in Fig. 4.

Consider the following sequence of features on an optical disk: Lm, Sl, Sm, Sl, and Lm. When the stylus is centered on the first Sl, as before about half the light falls on the land and slightly less than half the light falls on the adjacent marks. (Because of the gaussian profile of the optical stylus, a very small amount of light illuminates the second Sl.) Slightly more than 50% of the light is returned. If the stylus is centered on the next feature, Sm, half the light illuminates the mark and now slightly less than half illuminates the adjacent lands. Once again about 50% of the light is returned to the detector. Because this is above the 25% threshold, Sm is not detected. Adjusting the threshold will not help as the difference in light returned from Sm or Sl can easily be less than the noise.

For the third case, consider the sequence of features Lm, Ll, Lm, (L + S)l, and Lm. The notation (L + S) indicates that this feature's length is equal to the sum of the length of the long and short features. Because all the features are longer than the size of the optical stylus, it is always the case that either 100% of the light (for lands) or none of the light (for marks) is returned to the detector. The only difference between Ll and (L + S)l is that for the (L + S)l the location at which the signal crossed the threshold is shifted by the length of a small feature. This difference in length is easily detected.

The preceding discussion is the motivation behind modulation codes, which is the modification of the user's data to match the characteristics of the data storage "channel." Conceptually, the size of a mark recorded on the disk is reduced until it is the minimum that can be reliably detected when sandwiched between lands of the same length. The length of a feature is allowed to vary by the (much smaller) amount that can be reliably detected. In the limit the length is continuously variable; this corresponds to analog FM of the type used in video disks.

Another constraint must be considered in the construction of a modulation code. To measure the length of the features on a disk, a "ruler" of some sort must exist. A constant frequency square-wave signal of period equal to the smallest increment of feature length, called a clock, provides this function. However, an optical disk reader is a mechanical device; the speed of rotation of the disk may vary due to, e.g., imperfections in the ball bearings in the motor. The clock should speed up and slow down in tandem with the disk. A circuit called a phase-locked loop (PLL) will synchronize the clock to the data, but only if the PLL receives frequent updates. This is accomplished by designing the modulation code so that a maximum as well as a minimum feature length exists.

These points are illustrated by the modulation code used in the compact disc, where the pits in the substrate are produced with discrete lengths nT, where $n = \{3, 4, 5, ..., 11\}$ and T is the period of a fundamental "clock" used by the CD system. This is accomplished as follows:

1. The user's binary digital information is grouped into 8-bit bytes.
2. A look-up table is used to translate each byte into a sequence of 14 bits (ones or zeros). The translation table is set up so that the sequence of 14 bits never contains fewer than 2 nor more than 10 zeros between ones. This is called the "eight-to-fourteen modulation code" or EFM.
3. Pits are formed in the disk so that the edges correspond to the "ones" in the sequence.

Schematically, the function of the reading device is to sort the length of each pit scanned by the optical system into one of nine "bins", $3T$ through $11T$, to recreate the 14-bit sequence. Note that any individual pit can have a detected length between $(n - 1/2)T$ and $(n + 1/2)T$ and still be sorted into the correct bin. This gives the system a significant degree of noise immunity.

Occasionally severe noise will push the detected length of the pit into an adjacent bin, causing the pit to be incorrectly decoded. (In fact, the CD system is intentionally designed so that this will be a relatively common occurrence!) The principles of the previous paragraph can be extended to detect and to correct such errors. The bytes into which the pit lengths are translated can, of course, have any value between 0 and 257—they are the user's data. If a pit length is measured incorrectly, it will still be translated into one of those 258 values and there will be no way to detect that an error has occurred. However several of the user bytes can be gathered together and used as the input to an algorithm that generates new bytes, called "parity bytes." The parity bytes are then appended to the user's, forming what is called a "block" of data. The algorithm can be chosen so that (1) each collection of user bytes creates unique output bytes and (2) for all possible inputs, the algorithm produces only a small fraction of the total possible bytes.

On a CD, 8 parity bytes are appended to every 24 bytes of user data to form a block. Therefore there are $2^{8(24+8)} \approx 1.2 \times 10^{77}$ possible blocks, of which only $2^{8 \cdot 24} \approx 6.2 \times 10^{57}$ correspond to valid user data. If noise creates a random block, a less than 5 in 10^{20} chance exists that this would be a valid block and so it would be easily detected as an error.

However, noise does not create a random data block. The mostly likely event is that only 1 of the 32 bytes in the block are in error. If an invalid block is detected, then in principle one could search through the list of valid blocks, find the one that differs by only 1 byte, and use it to replace the erroneous block. Clearly searching through 10^{77} blocks is impractical. Fortunately, the algorithm for generating the parity bytes (the encoding algorithm) can be chosen so that a related algorithm (the decoding algorithm) when applied to the data block will generate (1) what is hopefully the user's data and (2) a "zero" if the block corresponds to valid user data or the location and correct value of the erroneous byte if it doesn't! This is called an "error correcting code" or ECC.

If the number of valid blocks is a sufficiently small percentage of the total number of blocks, it is possible to recreate the original data even if more than one byte is in error. On a CD, 24 user bytes are processed by an ECC encoder that appends 4 parity bytes. The resulting blocks are then processed by a second ECC encoder that appends the second 4 parity bits. Each of these two ECCs has the power to correct 2 erroneous bytes and detect, but not correct, 3 erroneous bytes. The specific ECC used, plus the two-step process, is called a "cross interleave Reed-Solomon" code (CIRC).[27]

While noise is likely to cause a pit to be detected as having the wrong length, resulting in a single byte error, a defect on the disk is likely to cause a "burst" of several bytes in error. On a CD, the ECC must be capable of dealing with not only noise and intrinsic defects, but also scratches, scuffs, and dirt caused by casual handling in the home environment. A CIRC is particularly well-suited to the situation of a combination of single bit random errors and burst errors.[27] On a CD, an additional step exists where the bytes in blocks output by the first ECC encoder are shuffled (interleaved) before being grouped and processed by the second ECC encoder. Each byte of a block from the first encoder ends up in one of 28 blocks, spread over 109 blocks, output by the second encoder. This means a defect, which without this scheme would have obliterated 24 consecutive bytes of user data, appears as a single byte error in 28 different blocks. In theory, the CIRC code employed on a CD can completely correct a defect 2.5 mm long. When CD technology was new, salesmen would drill millimeter-sized holes in a disc prior to playing it to demonstrate the system robustness achieved by this coding scheme!

Earlier it was mentioned that the CD system was designed so that errors are common. Although this is counterintuitive, it actually maximizes the capacity of the disc. Because of the small size of the features on optical disks and the state of current manufacturing technology, even under the best of circumstances defects cause approximately 1 bit in 10^6 to be in error. A rule of thumb for data integrity is that the error rate must be $<10^{-12}$, so some form of ECC is always required in optical disk systems. As the features on the disk are made smaller, the data capacity increases but eventually the resolution limit of the optical system is exceeded resulting in decreasing signal strength. Decreased signal-to-noise ratio causes random errors to increase. However, because one has already paid the price of including ECC, one may as well decrease the feature size (increasing capacity) until the random error rate is just below what can be handled by the ECC. On a CD, on the order of 1 block in 100 will be in error (≈10/s at standard "1×" play speed), and each will be corrected by the ECC.

Conclusion

While there is a certain poetry to the use of optical disks to store optical images, an analysis of the characteristics of both optical disk systems and digital image files leads to the conclusion that they are well matched. Even with clever encoding schemes, a high-quality digital image—particularly one intended for hard copy—will consume many megabytes of storage. Optical disk systems work most effectively with large, contiguous files.

Furthermore, the leverage of CD technology across many applications has resulted in this particular optical disk system being a very high-volume product with resulting low cost. About 90% of all personal computers are shipped with a CD-ROM reader. This large installed base makes CD the obvious choice for the distribution or sharing of digital images.

Eventually, DVD drives will be common components of desktop computers, with increased capacity and functionality over CD. Industry commitment to maintain compatibility with CD means that optical storage will continue to be a preferred method for the storage and distribution of digital images.

References

1. N. Negroponte, *Being Digital,* Alfred A. Knopf, New York, 1995, p. 10.

2. See for example, F. Haber, *An Introduction to Information and Communications Theory*, Addison-Wesley, Reading MA, 1974, pp. 14–22. The seminal papers in information theory are C.E. Shannon, A mathematical theory of communication, *Bell Sys. Tech. J.*, **27**, 379; **27**, 623.
3. E. Mayer and L. Von Salvini, On the evolution of photoreceptors and eyes, *Evol. Biol.*, **10**, 207 (1977). G. Halder, P. Callaerts, and W. J. Gehring, New perspectives on eye evolution, *Curr. Opin. Genet. Dev.* **5**, 602 (1995).
4. S. I. Tomarev, quoted in J. Travis, Eye opening gene, *Sci. News* **155**, 288 (1997).
5. R. Hoshino, Direct-overwrite removes MO application barriers, *Comp. Technol. Rev.*, **17**, 4 (1997).
6. T. K. Hatwar and D. G. Stinson, Control of exchange coupling for direct overwrite media using Pd metal interlayer, *J. Appl. Phys.* **70**, 6439 (1991).
7. D. Howe, Photo CD—Music for your eyes, *Opt. & Photonics News*, 58, (February 1993) and 58 (March 1993).
8. J. Larish, *Photo CD: Quality Photos at Your Fingertips*, Micro Publishing Press, Torrance CA, 1993.
9. A. Arai, Y. Shin, T. Ishiguro, and E. Hamada, Method of manufacturing an optical recording disc, U.S. Patent 5009818 (1991).
10. Y. Takagisi, K. Ootaguro, E. Hamada, and T. Fujii, Optical information recording medium and recording method, U.S. Patent 5080946 (1992).
11. E. Hamada, Y. Arai, Y. Shin, and T. Ishiguro, Optical information recording medium, U.S. Patent 5090009 (1992).
12. J. Foley, van Dam, Feiner, and Hughs, *Computer Graphics: Principles and Practice*, Addison-Wesley, Reading MA, 1992.
13. A. Marchant, *Optical Recording: A Technical Overview*, Addison-Wesley, Reading MA, 1990, p. 111.
14. A. Marchant, *Optical Recording: A Technical Overview*, Addison-Wesley, Reading MA, 1990, p. 113.
15. G. Bouwhuis, J. Braat, A. Huijser, J. Pasman, G. van Rosmalen, and K. S. Immink, *Principles of Optical Disc Systems,* Adam Hilger, Bristol, 1985. p. 125.
16. A. Marchant, *Optical Recording: A Technical Overview*, Addison-Wesley, Reading MA, 1990, p. 166.
17. D. Maydan, Micromachining and image recording on thin films by laser beams*, Bell Sys. Tech. J.* **50**, 1761 (1971).
18. P. Kivits, R. deBont, B. Jacobs, and P. Zalm, The hole formation process in tellurium layers for optical data storage, *Thin Sol. Films* **87**, 215 (1982).
19. D. Howe and J. Wrobel, Solvent-coated organic materials for high-density optical recording*, J. Vac. Sci. Technol.* **18**, 92 (1981).
20. D. Chen, J. Ready, and E. Bernal, MnBi thin films: physical properties and memory applications*, J. Appl. Phys.* **39**, 3916 (1968).
21. J. Biesterbos, Properties of amorphous rare-earth-transition metal thin films relevant to thermomagnetic recording, *J. Phys.* **40**, C5-274 (1979).
22. K. Watanabe, T. Oyama, Y. Aoki, N. Sato, and S. Wiyaoka, New optical recording material for direct-read-after-write (DRAW) disks, *Proc. SPIE* **383,** 191 (1983).
23. P. Kivits, R. deBont, and P. Zalm, Superheating of thin films for optical recording, *Appl. Phys.* **24,** 273 (1981).
24. G. Bouwhuis, J. Braat, A. Huijser, J. Pasman, G. van Rosmalen, and K. S. Immink, *Principles of Optical Disc Systems,* Adam Hilger, Bristol, 1985, p. 189.
25. A. Marchant, *Optical Recording: A Technical Overview*, Addison-Wesley, Reading MA, 1990, p. 327.
26. S. Golomb, Optical disc error correction, *BYTE*, 203 (May 1986).
27. W. W. Peterson and E. J. Weldon, *Error-Correcting Codes*, MIT Press, Cambridge MA, 1972, p. 371.