STORAGE AND COMPRESSION

Once images are captured and edited, they must be stored. Technically, there is no difference between storing digital images and any other type of digital data. The same magnetic disks used in automated library systems to store programs and bibliographic records can also be used to store images. But the capacity requirements may pose a problem for a system sized for bibliographic records because an uncompressed, monochrome image of an 8.5- by-11-inch page at 600 dots per inch requires about 2 Mb of storage; an uncompressed 24-bits-per-pixel color image of a similar-sized photograph requires 24 Mb. Even when compressed, image files require far more storage than is available on most automated library systems without a major upgrade.

The storage medium can be attached to a PC or a server; however, given that entry-level servers now cost under $3,500, it is wise to use a server that can support several concurrent users, rather than a PC. The proper place for a PC in an imaging program is as the image-capture workstation—the device into which the digital scanner or digital camera temporarily loads the images.

The amount of internal disk storage typically configured with a high-end Pentium PC (15GB) is not adequate for storing an image collection, not even a relatively small one. It is possible to purchase an upgrade to 75 Gb for some models at an additional cost of about $400, but that is still a small amount of storage for an image collection—and it allows only one user at a time to access the images.

An entry-level server from a company such as Compaq or Dell can accommodate from five to 24 concurrent users. It can be configured with two to four internal disk drives, typically 18.2 Gb SCSI-3 drives. Beyond that, external disk storage has to be purchased. An external 18.2 Gb SCSI-3 disk drive can be purchased for about $900. The best models have an average seek time of 5.7 ms (milliseconds or thousands of a second). This provides not only cost-effective additional storage for a small system but also good retrieval time.

Larger systems configured on mid-range servers—servers costing $10,000 to 30,000 and supporting 25 to 100 concurrent users—can accommodate larger disk drives; however, improving on the price and performance of 18.2 Gb drives is difficult. For example, a 36.4 Gb SCSI-3 drive has a seek time of 7.6 ms at a cost of about $1,800. You receive twice the storage at about twice the cost but suffer a one-third increase in retrieval time. Keep in mind that most mid-range servers can accommodate only a specified number of disk drives, so using smaller, faster drives can reduce the total potential disk storage capacity of the system. Most mid-range servers, however, can accommodate 80 to 120 Gb of disk storage using 18.2 Gb disk drives.

Disk drives should be backed up regularly. Not only is the backup required to restore files if a disk drive crashes, but it is needed at least once every five to seven years to refresh the database by copying it onto new media.

Alternative Storage Media

CD-R (Compact Disc-Recordable) has been a popular alternative to magnetic media because the disks cost less than $10 each and retain data for much longer than magnetic media—a minimum of 20 years and possibly as long as 100 years. Each CD-R has a capacity of 650 Mb. You can store up to 600 Mb on a hard drive, then burns a CD-R to clear the hard drive for new images. That leaves about 50 Mb

on the CD-R for operating functions. Multiple CD-Rs can be mounted in a jukebox with a capacity of as few as six and as many as 500 CD-Rs and four to six drives. A model storing 60 Gb of information (100 CD-Rs), however, can cost as much as $15,000. A similar amount of hard disk storage costs about $5,400. The other drawback is that a jukebox may take as long as 4.5 seconds to load a CD before data retrieval can begin.

Despite the longevity of information on CD-R, backing up the database is important.

DVD-R (Digital Video Disc-Recordable or Digital Versatile Disk-Recordable) is increasingly competing with CD-R because the capacity is as high as 8.7 Gb per DVD-R. Although the disks are a minimum of five times as costly as CD-R, their capacity is 14.5 times greater. The recording equipment is somewhat more expensive, but jukeboxes are not. The typical jukebox is configured with four to six drives. When a disk has to be retrieved from the jukebox and loaded into a drive, it can take as long as 4.5 seconds to load a DVD before data retrieval can begin.

A major drawback of DVD is that there is more than one format and not all formats are compatible. DVD, however, is backward compatible, meaning that a DVD drive can accommodate a CD-ROM, so you can add CD media to DVD jukeboxes.

DVD jukeboxes are available with capacities of up to 1.0 Tb (Terabyte). A jukebox of that capacity requires only a 3-by-4-foot floor area.

Again, backing up the database is important.

Magneto-optical (MO) technology, with 5.2 Gb double-sided capacity, combines elements of both laser and magnetic technology. Instead of using a magnetic read-write head, as in hard drives, MO uses a laser beam to change the temperature of the medium, which allows it to be written to by a magnetic field. MO drives are not subject to head crashes because the media and heads do not contact one another. Once new data is written to MO media, it can be stored for 30 to 50 years. The main drawback to MO technology is that it is still priced well above CD and DVD technologies.

Again, database backup is essential.

Backup

Many organizations provide backup to an image database by storing each image twice, either on two separate hard drives or a combination of hard drives and CD-R, DVD-R, or cartridge tape drive.

The best choice at this time for immediate availability of the data in case there is a disk crash is a hard disk RAID array, a cabinet of disk drives designed for writing information to two separate disk drives simultaneously. If the popular RAID 5 technology is used, 60 Gb of disk storage costs as little as $3,750; 300 Gb costs as little as $7,350. Although RAID storage is typically configured on a server, rather than a PC, it can be configured with a PC because both IDE and SCSI interfaces are included.

Another backup option is an 8mm cartridge tape drive. There are many models, but the most cost-effective store 60 Gb. The tape costs under $100 and the drive as little as $2,000. The main reason for using the medium only for backup is that the data transfer rate is a relatively slow 12 Mbps—twice the time required to retrieve data from a hard drive.

Whichever storage medium is chosen, the key to controlling the amount of disk storage is to use compression to store the images.

Compression

Given the massive storage requirements of image data, compression techniques become important. Also, it is best to store data and ship it over a network in compressed format, expanding it only at the point of use. Over the past few decades, an enormous amount of work has been done on compression algorithms for images. For still images, these can be divided into two categories: lossless and lossy compression schemes.

Lossless and Lossy Compression

A lossless compression scheme allows the precise reproduction of an image from its compressed form; a lossy scheme attempts to produce a similar image while throwing away some data. Lossy compression is more cost-effective as it can reduce an image to a much smaller size, but inherent in the economy is the potential for loss of detail. Each algorithm has its applications: lossless would be essential for detailed illustrations, and lossy would be appropriate for plain text.

The most widely used compression format is JPEG. It allows for a choice in level of compression. Although a compression ratio of 25:1 is common, the usual range is 10:1 to 40:1.

There is also a compression algorithm known as LZW. It offers 40% to 90% compression. It is a proprietary solution that is not widely supported.

CCITT or Huffman encoding, which was developed for facsimile transmission, is only occasionally used in digital imaging.

In the context of moving images, additional compression possibilities are introduced. Each frame can be compressed using either a lossless or lossy compression algorithm. In addition, interframe compression techniques can be used, perhaps sending only the changes from one frame to the next if only a limited number of objects in the picture move from one frame to the next. This would save substantial storage space. Interframe compression, however, does introduce some complexities. If a long sequence of compressed images has been stored using interframe compression and you want to see the 10,000th frame, all previous 9,999 frames must be scanned to compute the single desired frame. In random access applications you would typically introduce synch points from which the interframe compression algorithm would be restarted.

Image Database Management

Software for managing databases of images has lagged behind hardware storage technology. Most database management systems (DBMS) cannot store images, or, if they can, they store them as Binary Large Objects (BLOBs) rather than as an image data type. The DBMS cannot interpret the schematics of image retrieval.

The most widely used DBMS for imaging applications is Oracle. It has many extensions to support still images. Other DBMS products are also adding such extensions.

Storage of sequences of moving images in a DBMS is still largely unexplored and raises basic questions about the interface between a DBMS and client machines in a network environment.