I’m new to imaging. What should I know?

Imaging & Scanner Fundamentals

Data Connect Corporation

Delivering Better Workflows for Over 25 Years

DATA CONNECT

Image Capture vs. Text Capture

Scanners convert paper documents to electronic form by taking a picture of each page. The electronic image is composed of dot-like picture elements, or "pixels." To a computer application the image of a piece of text is not the same as the text itself. To a word processing program or a database, an image of the letter A is not the same as the letter A; it's just a pattern of dots. This means that one cannot search for the occurrence of a particular word or phrase in an image document; one cannot spell check it; one cannot edit the document – insert, delete, or rearrange the elements.

Fundamentally, a scanner's job is to create a high-quality image of each page – not machine-readable text – and send it to a computer. Converting an image to a text representation of the page uses a technology called optical character recognition (OCR). For handprint, the technology is often called intelligent character recognition, or ICR. Today, most OCR is provided in software as a post-process to scanning; it is not a function of the scanner itself, except in some very high-end special-purpose devices.

Pixel Depth

Scanners are digital devices. Each pixel in the scanner output data is represented by one or more bits, and the pixel's bit value is interpreted by the scanning software to mean black or white, a shade of gray, or a color. The more bits per pixel, the more colors or shades of gray can be represented in the output image, but with the tradeoff of larger file sizes and slower operation.

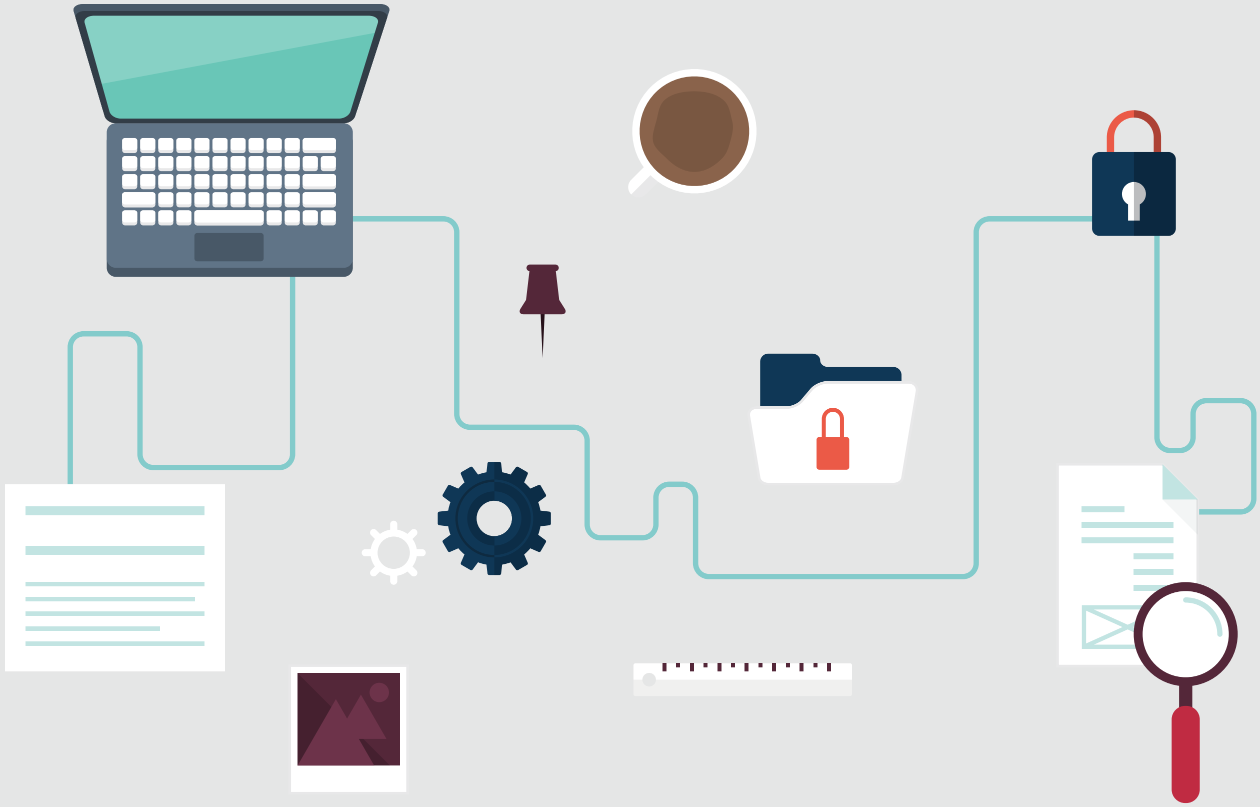
The output of scanners for business documents is typically bitonal, meaning only 1 bit per pixel, so a dot is either black or white. This reduction of bits in the output stream allows these devices to scan many pages per minute, and allows for much smaller file size than grayscale or color scanning.

Resolution

The number of dots per inch (dpi) of the original page is called the resolution. The higher the resolution, the higher the image quality, again at the price of larger file size. A web page graphic may call for scanning at 100 dpi or less; high-end graphic arts flatbed scanners have resolutions ranging from 5,000 to 10,000 dpi.

Speed and Paper Handling

Increasing a scanner's pixel depth or resolution does not dramatically increase its cost, but increasing its speed and paper-feeding capabilities does. Speed is measured in pages-per-minute (ppm), or in the case of duplex scanners, the images-per-minute (ipm) of front and back sides.



Production

The term "production scanning" refers to high-volume capture, usually of letter or legal-size business documents or records, for workflow processing and image archiving. Examples include capture of insurance claims, loan applications, invoices, and tax returns. Production scanners often have dedicated operators, and clerical staff to prepare documents for scanning by removing staples and inserting batch separator sheets. These scanners also must be able to scan over a thousand pages per day for several years without wearing out.

Today, most production scanners are sheet-fed devices that run at 200 or 300 dpi resolution and output 1 bit per pixel. Speeds range from around 40 to over 150 ppm, and the devices are designed for high daily volumes. These scanners run too fast to connect to a PC through a standard parallel or serial port. Instead, they are attached either through a special high-speed video interface card, or through a SCSI card. The cost of these devices is anywhere from $10,000 to $100,000.

Exception

Most production scanning environments also have a need for exception scanning to handle documents that could not be handled by the production scanner, possibly because they were bound or oversize, torn or fragile, or rejected for image quality. Examples include patient records, delivery receipts, health claims, tax returns, and scanning of old paper archives.

Since these documents are often handled one at a time, speed is not as big a factor as convenience and flexibility. Flatbed scanners, particularly those with an attached auto-feeder and designed for high volume, are commonly used for exception scanning. They also typically run at 200-300 dpi, 1 bit output, and connect via SCSI or video interface cards. The cost of these devices is typically $5,000 to $15,000.

Workgroup

In a workgroup scanning environment, a moderate daily volume – perhaps 50 to 1000 pages per day – must be captured, either for a business application or for routine office functions. Examples include correspondence management, patient records, resume tracking, and low-volume implementations of production scanning applications.

To handle a wide range of document types without the need for a separate exception scanning, workgroup scanning usually calls for a high-speed flatbed scanner with an attached document feeder. These usually run at 20-50 ppm, scan at 200-300 dpi with 1 bit per pixel, and include advanced image processing.

Some workgroup scanners are being offered as walk-up network appliances, operated from a copier-like control panel instead of a computer keyboard. The cost of these devices runs from $3,000 to $10,000.

Document Transport

Most scanners operate by moving the paper tray past a stationary sensor/optics assembly, a function provided by the document transport. (In a flatbed scanner without a document feeder, the paper does not move, so there is no transport component.) The transport is responsible for moving the paper quickly and reliably – i.e. without jamming – from the input tray to the output tray.

The feeder is the most critical part of the transport components. The ability to handle mixed paper widths and weights without jams or double-feeds, and to produce minimal skew in the scanned image, are functions of the feeder. Manual feeders accept pages inserted in the scanner input slot. Even some production scanners sometimes use manual feeders to eliminate occasional jams and double-feeds in batches with mixed size and weight paper.

An auto feeder (sometimes called an auto document feeder, or ADF) allows stacks of 30 to 500 pages or more to be placed in the input tray and fed automatically, without human intervention.

Document Transport, continued

Obviously the quality of the paper to be fed presents limitations to the paper-handling mechanism. Crumpled paper, torn edges, staples, rivets, paper clips, glue, etc., present problems for any automatic handling mechanism, and therefore the scanned documents must be prepared in advance, and tested for use with the ADF.

Imaging Assembly

The imaging assembly includes the components responsible for converting marks and color shades on the paper into electronic data. At the core of the imaging assembly is the image sensor. Most scanners use a linear Charge Coupled Device (CCD) for the sensor. This type of CCD is a linear array of several thousand light-sensitive elements in a single integrated circuit.

Light reflected from the page passing by the imaging assembly is focused on the CCD, and each element accumulates an electrical charge in proportion to the light it receives. White parts of the page will generate more charge than black parts of the page.

Imaging Assembly, continued

Then the CCD acts like an analog shift register, with each element transferring its accumulated charge to the pixel adjoining it, like a firefighter’s bucket brigade, with the charge from the last picture element transferred out of the CCD into the processing electronics. This entire process is repeated for every line scanned on the page. For a 300 dpi scanner running at 60 ppm, this means 3300 lines are imaged like this in less than a second!

In addition to the CCD itself, the imaging assembly includes a bright source of illumination, plus lenses and mirrors to direct the light onto the paper and to focus the reflected light onto the CCD. Most scanners today use a high intensity fluorescent lamp that shines a bright line of illumination on the paper.

In flatbed scanners without an ADF, the entire imaging assembly, including illumination source, optics, and sensor, moves while the paper lies flat on the glass plate. In sheet-fed scanners and when using the ADF of a flatbed scanner, the imaging assembly is fixed while the paper moves. Duplex scanners include two separate assemblies to image the front and back sides of a page.

Imaging Assembly, continued

A CCD designed to capture color information includes three linear sensor arrays with red, green, and blue filters respectively, allowing full color information to be captured in a single pass. A CCD designed to capture grayscale or black and white has only one array and no built-in filter. Grayscale and monochrome scanners often use green lamps, while color scanners use white lamps.

Speed

Speed is critical in production scanning, and a scanner’s speed largely determines its price. A common misconception about scanning throughput is to estimate daily volume by simply multiplying the scanner’s rated speed in ppm by the number of minutes in an eight hour day. Using such a calculation, a 60 ppm scanner would be expected to produce 3,600 pages per hour, or close to 29,000 per shift. When you include time to move batches in and out, remove jams or misfeeds, and make other operator adjustments, the actual maximum daily throughput may be only half that or less. In practice, scanning more than 2,000 pages per hour requires a streamlined document preparation (doc prep) operation and trained operators.

Note: *Production scanners typically specify their speeds at 200 dpi; if you want to scan at 300 dpi, the speed may be slower.*

>

Finding the Ideal Scanner Speed

Scanners rated at 60 ppm or faster are used in centralized mailroom operations at banks and financial institutions, insurance companies, utilities, transportation companies, retailers, and government agencies. Applications include loan applications, insurance claims, accounts payable, shipping documents, tax returns, and other government filings. They are also used by service bureaus for applications such as litigation support and back file conversion of any document repository.

Scanners rated around 40 ppm are most often used in production imaging applications where scanning is distributed in user departments, and occurs in bursts of activity of an hour or two each day. Applications include correspondence management, insurance underwriting or claims investigation, or dispute resolution.

Scanners in the 20 to 25 ppm range are typically flatbed designs with ADF, used for exception scanning or small batches of document pages. They are also used in shared network scanners for workgroup filing, fax, and graphics capture.

Finding the Ideal Scnner Speed, continued

Scanners with speeds rated under 10ppm are able to connect to a PC through a standard parallel or serial port, and are used in applications that scan one or a few pages at a time, such as graphics creation, personal filing, or desktop fax.

Flatbed vs. Sheet-Fed

If you only need to quickly scan batches of cut sheet paper, particularly business documents, a production sheet fed scanner is recommended. In general they offer faster transports and take up less desktop space than flatbeds, but typically output binary or grayscale data only, not color.

If you need to scan single pages in color and high-resolution and want to accommodate oversize or bound material, a flatbed scanner is recommended. If you want a scanner primarily for business documents, but you also want the ability to scan color, bound, and oversize material, and can sacrifice a little speed, a flatbed with an integrated ADF is your best choice.

Finding the Ideal Transport Type

Sheet fed scanners are used in most production scanning applications, including loans, claims, accounts payable, records management, as well as workgroup fax and filing.

Flatbed scanners are used in most graphics capture applications, particularly when just a portion of a page is scanned, and for OCR capture of books and magazines. They are also used (with ADF) when a one-size-fits-all device is desired.

Duplex vs. Simplex

Some documents to be scanned are printed on both sides. Duplex scanners can scan front and back in one pass, and provide the fastest and easiest solution to this problem, but are more expensive than single-sided, or simplex, scanners. Until recently, duplex was only available on sheet fed scanners, but it is now available in the ADF of flatbeds as well.

Software duplex is an option if the volume of duplex scanning is relatively small. With software duplex, a stack of pages is placed in the ADF of a simplex scanner, and when all the pages have run through the scanner, the stack is flipped over and the back sides scanned. Software correctly reorders the pages in the output image file.

Applications That Require Duplex

Applications of duplex scanning are high-volume forms processing, such as tax forms, checks, contracts, and government records. Duplex scanners are also required for litigation support, contracts, and OCR conversion of books and manuscripts. If only occasional documents or folders are printed on both sides, software duplex works well here.

Paper Handling

In production applications, paper handling is probably the most critical scanner issue. The ability to handle creased, curled, and torn originals without jamming saves major expense in doc prep operations. The ability to handle a wide range of paper weights and finishes, from flimsy carbonless forms and onionskin to thick card stock is important, and even better is the ability to mix weights in a stack and auto-feed them without jamming.

You shouldconsider also the size of the pages to be scanned. Sheet-fed scanners that can accommodate 11 inches or wider cost more than those that can only can 8.5 inch widths.

Paper Handling, continued

Most auto-feeders have a minimum width that they can handle. Auto-feeding mixed document widths in a batch is a common difficulty for scanners today. If this is a requirement in your application, look for scanners that claim they can do it well.

Skew is a common problem in production scanning environments, particularly with documents of mixed widths or weights. Skewing increases compressed file size and degrades OCR accuracy, and while software can de-skew the image output, it's better to eliminate it in the scanner itself.

Double-feeding is a constant worry in batch scanning with ADFs. When two pages are fed into the scanner at once, some image data is not captured. Some scanners include sensors that detect double-feeds and pause for operator attention.

The size of the input tray is an important paper-handling feature. Beyond 500 sheets, the weight of the paper in the stack becomes a technical issue. Some production scanners can hold 1000 pages or more in the ADF.

Applications That Require Mixed Paper

Production scanning applications with mixed paper size and weight include a variety of those in which business forms are submitted along with attachments, such as insurance claims, loan applications, and tax filings.

Another set of applications includes those where a variety of small-format documents such as receipts, remittances, credit card slips, or shipping documents must be scanned. In some applications such as tax filings, envelopes with postmarks must be scanned as proof of timeliness. Multipart forms are still used in insurance, healthcare, transportation, and government applications, and this lightweight paper is often mixed in with heavier weight paper attachments in scan batches.

Duty Cycle

Just because a scanner is rated at 25 ppm doesn't mean it was designed to scan thousands of pages per day, every day. If you plan to average over 500 pages per day, you need to consider the duty cycle of the scanner, usually specified as a number of pages per month, or a total machine life specified in pages. For production scanners, sheet-fed devices typically have higher duty cycles than flatbeds with ADF.

Matching the Duty Cycle to the Application

A wide variety of applications are used at average daily volumes of over 500 pages per day, but duty cycle information is not typically printed on a scanner's data sheet. For over 2,000 pages per day, a production sheet-fed scanner is usually recommended. Between 500 and 2,000 pages per day, some flatbeds may be rugged enough, but review the warranty carefully.

Image Processing

All scanners offer basic control over brightness and contrast. Many production scanners provide advanced dynamic threshold, including the ability to follow changing page backgrounds, read light pencil and blue pen markings, and generally produce a good image from a wide range of poor quality documents using a single default setting. In some cases this functionality requires a special image processing option in the scanner.

Artifacts, such as dirt and smudges are frequently introduced to documents when they are handled. The most common artifacts can be introduced during the scanning process by dirt or dust in the scanning mechanism.

Noise can be particularly confusing for OCR engines that are attempting to read characters from the image and may attempt to recognize noise as part of a character it is trying to read. Noise removal enhancement will automatically remove "stray" particles from the image. Sometimes this is provided inside the scanner, sometimes on the scanner interface card, and sometimes in software on the host PC.

Image Processing, continued

Image processing can also automatically discriminate text and continuous tone portions of a document so that the appropriate threshold technique can be applied.

Applications With Dynamic Thresholds

While all bitonal scanning benefits from dynamic thresholds, applications where it is critical are those where users have filled out forms by hand in pen or pencil, particularly when the form is on colored paper or has shaded areas. Examples include direct mail replies, tax filings, government records, shipping and receiving documents, remittances, and loan applications.

Dropout Lamps

Many production scanners allow special colored lamps to be used in the image assembly for forms dropout. Lamp colors must be matched to the color of the dropout ink on the form.

Applications That Require Dropout Lamps

Dropout lamps are mainly used in centralized processing of large volumes of a single type of form, such as 1040EZ tax forms, HCFA-1500 health claim forms, subscriptions, and direct mail response forms.

Endorser

Some production scanners offer an optional endorser—a device that prints an ID number on documents as they enter or exit the scanner. This is useful for connecting the original paper to an indexed image. The scanning room is often a bustling environment filled to the brim with documents either being prepared for scanning or removed after scanning.

Endorser, continued

Obviously a mix-up between scanned and un-scanned documents can cause considerable problems. For this reason many scanners include the ability to print a verification (endorsement) on the document to indicate that it has already made one pass through the scanner.

Applications for Endorsers

Endorsers are generally used when the original paper is used for certain processing or is maintained as a legal record even after scanning. Most common applications involve financial transactions and government records.

Some regulatory authorities, such as the Securities Exchange Commission, and legal precedents require that certain original documents be retained for a specified period. The ID number printed on the page usually links the image database record with the physical batch and indirectly to the archival storage location of the original paper.

Barcode Patch / Code Reader

Barcodes can be read very reliably inside the scanner using an auxiliary device, and are commonly employed for top-level indexing purposes. A crude kind of barcode called a patch code is also employed on separator sheets inserted at doc prep time between documents in a batch, or sections in a document. An optional barcode or patch code reader in the scanner can be used to route specific documents or sections for indexing or special processing.

Applications for Barcodes

Barcodes are useful in forms applications because the barcode can be pre-printed on the form. Examples include shipping documents, insurance claims, and tax forms.

Applications for Patch Codes

Patch codes are only used in high-speed scanners (over 60 ppm), on applications where variable length folders or filings are scanned together as part of a batch. Examples include health claims, patient billing, tax or UCC filings, and mortgage loan applications.

Maintenance

The trend today is for customer-accessible maintenance, rather than requiring a field service call. The most common maintenance activities are lamp replacement and transport cleaning.

Many scanners today make it very easy to replace a lamp without special alignment tools, and to recalibrate the imaging assembly. Similarly, the current generation of scanners has been largely re-engineered to make clearing jams and cleaning the feed rollers quick and user-friendly.

Scanner selection should consider these two basic operations in addition to speeds and features.

Host Interface

Scanners connect to their host computers in a variety of ways. You should consider both the hardware and software interfaces.

From a hardware standpoint, there are three types of interfaces for production scanners. A video interface transmits uncompressed image data to a proprietary scanner interface card, where it is compressed without slowing down the scanner. Video interface cards also perform advanced image processing functions such as image rotation, de-skewing, noise removal, black border removal, and barcode reading.

A SCSI I interface uses a relatively inexpensive industry-standard interface card. With SCSI, often images are compressed at high speed on the scanner, and any image processing performed in software rather than on the SCSI card.

Network interfaces are starting to emerge in some scanners, particularly those used in a shared workgroup environment. These essentially embed intelligence inside the scanner, allowing the scanner to attach directly to an Ethernet just like a network printer.

Host Interface, continued

There are three major kinds of software interfaces.

ISIS is a widely used software interface that supports almost all production scanners. An application written to support ISIS can automatically use any scanner with an ISIS driver.

TWAIN is a standard interface for personal/casual scanners and for graphic arts devices. Most common desktop applications that support scanners do so via TWAIN. A new version of TWAIN in development will support high speed scanners also, as an alternative to ISIS.

Scanner card APIs (application program interfaces) and ActiveX Controls are proprietary software interfaces specific to a particular vendor's scanner card. (An ActiveX Control, also called OCX, is a high-level API that includes a user interface, used with programming tools like Visual Basic.) An application written to such an interface can take advantage of special features on the card, and support all scanners that the card supports, but the API or control cannot be used without the interface card. Some video interfaces cards support ISIS and TWAIN in addition to their own API.

>

Working With Forms

A significant portion of high-volume scanning applications requires capturing data from forms for workflow processing or records retention. To save the labor of full key entry, a number of image processing technologies can be integrated in a pipelined capture sub-system to do a big part of this automatically. These include:

Scanning – The scanner should include an ADF with a high-capacity input tray, barcode and patch code recognition, and possibly a dropout lamp.

Image Enhancement – Image rotation, de-skewing, and noise removal should be automatic.

Quality Assurance (QA) / Rescan – Automatic QA from some vendors can detect whether an image was scanned at the proper brightness and contrast, or excessively skewed or cropped, and reject the page for exception scanning.

Form Identification – Software can identify the particular form, matching it to the signature of a blank form recorded in a database.

OCR/ICR – Once a form has been identified, the appropriate OCR or ICR engine can be applied to each zone (field) on the form to recognize the data.

Working With Forms, continued

Multiple recognition engines may be used simultaneously to increase throughput and reduce errors. Characters that fail to achieve a specified confidence level in the recognition process are marked as rejected.

Reject Reentry – Fields with rejected characters are routed to key entry operators for manual entry. Reject reentry software is highly optimized for high key-from-image speeds.

Validation – Recognized field data can be tested against specified validation rules, including lookups of account information, postal address data, etc. Data that cannot be corrected automatically is rerouted for reject reentry.

Export – Images and validated form data must be formatted according to the requirements of the document management system and business application that will use them.

Most production forms-reading installations include these functions. In many cases, custom software is required, but increasingly the trend is toward integrated capture subsystems that provide a "productized" framework for forms reading and faster, less costly deployment.

Questions? We’d love to hear from you.

Phone: (303) 840-7477

Email: tech@dataconnectcorp.com

Web: www.dataconnectcorp.com

Office: 6555 S. Kenton St., Suite 310
Centennial, CO 80111

DATA CONNECT