Scanning Archival Material

By
Daniel D. Whitney

As we move from the 20th Century into the information age of the 21st, mountains of important material are being lost because there no way to index or even know where it might be. Established archives hold only a small fraction of the documentation relevant to the historical context of many exciting topics, with much scattered in personal collections and corporate files. The tragic stories of how much valuable material has gone into the dumpsters bring tears to the eyes of historians. Still, those who do seek out material from history's dark corners are faced with an information dilemma. How to collect the material in an efficient and useable manner for study and analysis at a later time. For the past 30-40 years we have had the ability to "Xerox" or photocopy documents, usually at relatively poor quality, but necessary given the usually brief opportunities for access and collection. Today we have new technologies available that not only allow acquiring copies of documents at a higher standard of quality, but even recovering data from documents that themselves are fragile or of such poor quality that they cannot be Xeroxed. These methods use optical scanners connected to a computer. The purpose of this article is to provide researchers with some guidance toward getting the most from their equipment, while minimizing the collection time and sizes of data files.

The first decision to be made when scanning documents is to determine the purpose or use to be made of the collected information. This will determine the size of the resulting computer data file and define the uses to which it can later be put. For example, a Xerox like quick copy can be made using "gray scale" (256 gray colors) at 100 dpi quite rapidly, but the resulting file will not be useable by an Optical Character Reader (OCR works best with two-color [B&W] images at 300 dpi).

If the purpose is to just have a copy of a document then it is probably most expeditious to simply Xerox it. If it is intended to establish an electronically searchable index to raw data, then a scan is in order. If it is possible that at some future point the information might be reassembled into an electronically searchable text document then the nature of the original scan is somewhat defined.

Scanners are marvelous pieces of equipment. Today you can buy one with outstanding features for under $50.00, capabilities that would have cost over $2,500 just 5-6 years ago. While each scanner usually comes with the software drivers and controls needed to operate it, there can be a wide range of control features available in the different units. "One-Button" scanners are not the best for the kind of work we are going to be doing. We need to be able to control and adjust the various parameters in order to get the kind of results we want.

How Scanners Work

Scanning paper documents can be best done with a "flat bed" scanner, in fact this is the only type allowed by the National Archives and others because of the possibility of damage by document feeders. In these the document is placed face down and the scanning head and light source are moved on rails underneath. The scanner head contains several rows of miniature light detectors, and for a "color" scanner, there will be detectors sensitive to red, green, and blue light. In this way determining the relative amounts of R-G-B at each point on the document allows digitizing the full color spectrum. Most non-commercial scanners today provide an "optical" or mechanical resolution of 600 dpi, that is, 600 dots per inch along the scanning head. Some will offer higher resolution via "interpolation", where software in the scanner mathematically provides data values for the interval between "dots". A digital device does not see "color", rather each of the three primary R-G-B channels simply determined the amount of "color" it is seeing. These quantities are then divided into 256 levels. While this may not seem like very many shades of color, it has been determined that with respect to "gray" (on the black-white scale) humans can only discriminate about 80 shades, so 256 is pretty good. There is also a digital reason for this number. Each value in a computer is described by "bits" of data arranged as "bytes", or "words". If one "word" (byte) is committed to remembering the value of a color seen by the scanner head from one of its 600 dots at a particular location in the scan, then we need to reserve enough memory to contain all of the possible values that that dot can have. This is 256, and can be represented in digital space (where any particular memory location can have only one of two values, a "1" or a "0") by 2^8, that is by eight data "bits".

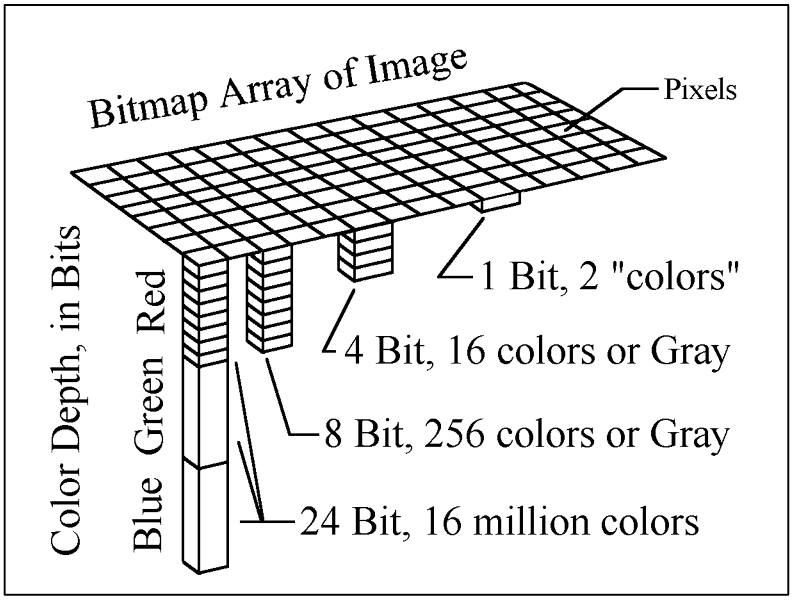
At a scan resolution of 600 dpi each one square inch of document can be described as requiring 600x600 or 36,000 dots, called "pixels" when in an array. If this were our image, when saved as a computer file it would require 36 KB (kilobits) of memory. This is the value that you see when you list computer files and their "size" is reported and is the type of file known as a "bit map". If the image was an entire 8-1/2"x11" page, at 600 dpi it will produce a file 3,366 KB, one for each "pixel" in the array, for a total of 3.4 MB! And this is for a file in only black and white. While old-timers might think this is a shocking waste of computer memory, with modern processors and storage media high-resolution files are cheap compared to the trouble of re-accessing the original document.

Another issue is to understand the effect of scan resolution (dpi) and printing resolution. Many of us have printers that print 720 or even 1440 dpi. This is "dots per inch" and is the number of R-G-B dots the printer is capable of producing. Since it takes a set of R-G-B dots to become a "pixel", the actual number of pixels per inch is really one-third the number of dots. Conversely, the scanner is actually limited by the mechanical (optical) spacing between its sensors, which are producing a set of data for each "pixel". This means that a printer at 720 dpi would need a scan done at 240 dpi to achieve full resolution. Any scan done at higher values will have the extra data mathematically reduced by the printer driver during the print process. For high quality magazine printing the standard is 133 "lines per inch". While not quite equivalent to dpi the effect is the same. Standard practice in the mass publishing business is to scan at 1.5 times the resolution of the printer. If you want your pictures for books and periodicals to "snap" the scans should be at a minimum of 600 dpi. For archival purposes, when you are scanning a quality image or drawing using a high resolution to provide a good basis that anticipates future needs, and developments in printing equipment. The real advantage of having high resolution scan capabilities is that small images can be enlarged in the final print without degradation in quality. For line-art or drawings, using high resolution is necessary to minimize the "jaggies" or stair-steps seen in otherwise straight lines. High-resolution scans of line-art may not cause excessive file sizes if compression formats like LZW-TIF are used. They will require your computer to have a fairly large amount of RAM memory in order to process or manipulate the image.

With a color scan each of the three R-G-B channels produces a 2^8 word for each pixel, or point on the scan. Our 8-1/2"x11" color scan at 600 dpi will result in a file some 10.098 MB in size, this is very large and as such will require a lot of computer processing time and memory to store and manipulate. Scanners described as offering "24 bit" color will use this format. Some newer scanners are described as doing "30, 32, 42, and even 48 bit" scans. These "ratings" describe the internal data processing within the scanner, almost all scanners in fact output their files in the 24-bit format. The actual amount of "color" data is not improved over the original 256 bit level, these scanners can provide a better color image because they incorporate a control feature known as the "gamma correction". This has the effect of allowing the operator to bias the data collection so that the optimum set of 24 bits is used in the final image file. This is particularly useful when it is desired to pull information out of dark prints or shadow areas. Most programs that use scanned data can only operate on 24-bit files, although the Adobe *Photoshop* program will handle 48-bit files.

Another aspect of a color scan is the "number of colors" contained in the image. "Full Color" is often shown as "16 million colors", but how can this be? Well it is the product of 256x256x256, equating to 16,777,216 possible hues and shades of color that can be represented by the "color depth" provided by one 24-bit word representing each R-G-B pixel in a scan.

Obviously scanning full page documents at full color results in very large files, and all sorts of other problems, not the least of which is that it may not be the best representation of the original document. This is a case where "more" is not necessarily "better". Look at Figures 1 and 2, which show an aged and discolored blue-line document. Figure 1 is full color, while Figure 2 is in 16 colors. With a million times fewer colors, the two documents still look alike. This would not be true if the image were of a color photograph.



A major decision is whether or not to scan in color. For full color originals, such as color photographs, the choice is probably obvious, set the scanner for "16 million" colors. However, if the document is a color schematic for an engine oil system it is appropriate to scan at "16 colors". Sixteen colors ($16 = 2^4$) can be represented by a bite (word) using only four bits of memory. In this case the three channels produced their usual 24 bits of data during the scan, but the scanner processed the data to resolve the 16 million colors down to 16. The result is a file requiring only one-sixth the amount of memory. A similar approach is used when either 256 color or the gray scales are used. In these cases eight bits are used for each pixel and the resulting file is one-third the size of a file at the same resolution, but retaining the "16 million" colors.

Even more savings can be achieved if the original document is a black and white text or drawing. Here the scan is done with a color depth of "1-bit", which represents two "colors", black and white. In this case the pixel count equates to the size of the bitmap file.

In summary, the scanning resolution (dpi) determines the number of "pixels" in the image. The "color depth" sets the size of the word necessary to contain the information at each pixel, that is, whether it is black or white (1-bit), black and white (256 grays, 8-bit) or normal color (usually 256 colors, 8-bit), or full color (24-bit). The combinations of selected resolution and color depth determine the size of files, and importantly, how the image can be used. When images are to be published in color it will be necessary to convert the R-G-B files into CMYK color format. Any of the good photo editing programs can output your files in this separated way, making them useable by the publisher. Black and White images should never be scanned in "color", rather scan as 256 shades of gray. These are then directly useable by the publisher.

Using scanned document files

With an appropriate scan and a software program with an OCR (Optical Character Recognition) capability, it is possible to convert the scan bitmap into a text document, such as a Microsoft Word "doc" file. As a text file it can be searched, formatted, edited, and saved in file formats that are considerably reduced compared to those from a bitmap. Applying this software is somewhat time consuming, but may be justified for some historical documents that are otherwise being lost due to age, rarity, or cost. Examples include aircraft manuals where their historical or technical content justifies the effort, for otherwise the material is not available in a usable format. The additional benefit of having a document that is fully searchable allows quick retrieval of otherwise obscure information.

Most OCR programs (*OmniPage Pro*, *FineReader Pro*, *Readiris Pro*) work best with original documents scanned at 300 dpi. A document scanned at 300 dpi will be only one-forth the size as one done at 600 dpi, assuming that both are done at the same color depth. Color depth for a document to be processed by OCR should also be done as a B&W drawing, the 2-bit setting. This gives the OCR program the cleanest image to process. If the scan is done on the gray-scale (256 grays), and then an image processing program is used to reduce the image to "2-bit", you will find that the OCR process will have a higher rate of incorrect recognition's. Spell-check can correct a number of these problems, but it can be a real mess when dealing with numbers. This is an example of where the scan needs to be done with the final use in mind. If there are B&W photos on the original document, and it is intended to OCR, it is often best to do two scans of the original, one a 300 dpi/2-bit of the entire page, and the second a 600 dpi/256-gray scan of the photos. The two are then recombined in the final processed document.

Exploiting the Advantages of Scanning

Not every original document is ideal for copying. Look at Xerox's made of color originals, or originals that are on flimsy carbon paper, documents on colored or discolored paper, or worse yet, thin paper printed on both sides. A scanner can provide excellent results in each of these instances, though it may require a little effort at the time. A lesson to remember, is make these types of corrections at the time of the scanning, do not try to use an image correction program after the fact. They can help, but it's not the best.

Most scanner operating software allows the resolution (dpi) to be selected as well as the color depth. When faced with the objective of getting an OCR quality scan from any of the above troubled originals, the scanner "threshold" control feature is of great help. Here you can select which of the three color channels is going to be used to obtain the "black & white" information and at the same time the degree of saturation. See Figures 3, 4, and 5 for examples. On a thin page, which allows the back side to show through, this can be set so that only the front image is recorded, and that in a desirable high contrast B&W. By trying the different color channels it is possible to compensate for discolored paper, and even do color scans with the result that a yellow background from the paper will come out "white", and with fully saturated colors. Just experiment.

File Compression

While a "bitmap" can accurately describe the color and textual content of an original document it does result in very large files. Fortunately there are "compression" algorithms (processes) available that will reduce size of the file. These algorithms are of two types, "lossy" and "non-lossy". As the names imply, compression can require compromise.

There are many different approaches to compression, such as recognizing that if a scan line is seeing mostly white, with an occasional series of black pixels, then converting the bitmap into a format that describes the locations of where white switches to black and vice-versa. This may considerably reduce the size of the file, and not lose any information in the process. While this may work well on a B&W drawing, it may result in an increase in file size when applied to a page of typed text.

Other approaches average pixels in an area together and then reconstruct by considering the similar content of adjacent areas. These approaches considerably reduce the amount of information in the resulting file, and are known as "lossy" formats. The GIF and JPG formats used extensively on the internet are examples. Since most computer monitors are only able to display 72 dpi images the losses are not usually noticeable. However, if a typical image is downloaded from a webpage, and then printed, the results are often unusable. The table at the end of this article gives particulars on several of the commonly used file formats.

Within most of these compression algorithms it is possible to control the amount of compression, and hence the quality (and file size) of the image. Repeated saving of an image can further degrade its quality. Additionally, decreasing the degree of compression of an image that has already been compressed will not improve its quality. Likewise, if an image scanned at 300 dpi is processed with a program like *Corel Draw*, *Paint Shop Pro,* or *Adobe PhotoShop* and increased to 600 dpi all that is changed is the size, not the quality. A print of the enlarged image will quickly show the degradation. The only way to improve the image quality is to rescan the original at a higher resolution.

Adobe Acrobat Format

The Adobe Corporation has provided a software product that has become the defacto standard for archiving and exchanging large files containing text, data, and images. While they market the program that creates these ".pdf" files, the Acrobat Reader is freely available. When material developed in a wide range of other word processing, analysis, and image programs is "printed" through the Acrobat software the resulting .pdf files can be viewed, searched, and printed by anyone with the free reader. The reader can be included on a CD-ROM with the archive material meaning than any future user of the material will have available the reader, no matter what the form of the technology of the time. Since scanning of documents, no matter how efficiently done results in many large files, and even after OCR and other processing, the final documents can still be large. This recommends the CD-ROM method for storage and circulation of the final documents. In many cases the original scans can be included as backup material, along with the final. By way of example, the entire operations, maintenance, overhaul, and erection manuals for the P-51A were placed on a single CD. The resulting .pdf is searchable, printable, and formatted to look like the original manuals.

Transparencies

Slides, negatives, and other transparencies can also be scanned. There is a difference in the equipment required as the light source must be transmitted through the transparency rather than "reflected". Some flatbed scanners have a top cover that provides the needed light source. These are quite suitable for large negatives and the like, but the results will not be satisfactory for 35mm slides. The problem is the mechanical limitations of the 600 dpi scan head, there are too few pixels. Specialized slide scanners are available, with optical scan capabilities of 2,400 dpi and above. These scanners can produce 50 MB files that are suitable for making prints as large as 11"x17". Commercial scanning services are available and can handle slides and other unique transparencies. Their equipment is of a higher standard than that usually used by the public and business markets, but unless the final product is to be a magazine cover, calendar, or other special use there is no need to seek out commercial class equipment. When it gets printed in a magazine the resolution is about 150 dpi at the best.

So How Should I Scan and Process Documents?

As stated above, first decide how and for what the scan is to be used. If it is a simple B&W document, typed or drawn, then do a simple 2-color, 1-bit scan at a resolution of at least 300 dpi. This allows it to be OCR'd, and will print satisfactorily on today's 600 dpi class printers.

If it is a B&W photo, then scan as 256 grays. Do not scan as color, for the files will be three times as large and the image will probably not be as good. Photos can often be scanned at 300 dpi quite nicely, particularly if the original is an 8"x10", for the final image when printed in a book or magazine is usually going to me much smaller, and the file will certainly be. If the original is a 2"x3" snapshot, then scanning at 600 dpi will result in an image that can be nicely used in a 4"-6" size range. This is also where the fully-interpolated resolution capability of the scanner might be useful.

Scans of printed color plates (not photographs) are usually best scanned at 256 colors, rather than 16-million. This is particularly true of color drawings and diagrams. Depending on ultimate usage the resolution should be as large as is convenient in terms of scan time and file size. Interestingly, scanning a full page color drawing at 600 dpi may in fact produce a bitmap file of nearly 34 million pixels, the actual file, if saved as a compressed TIF may be in the range of 2-3 MB. The reduction is due to the amount of white space. When you reopen the file it will still have the entire 34 MB worth of pixels, presented at the selected color depth. If the drawing has only a few colors, then be sure to set the color depth to 16.

Full color photographs will consume a lot of memory. The only practical control is on the scanning resolution, so keep in mind how large you might want the final image.

Issues and Integrity

With a scanned document you can do a lot of things, both good and bad. Using OCR and digital techniques text and even photographs can be manipulated, becoming what they never were before. Since it is possible to scan almost any document or image, it is critical that we honor copyright laws and the content of the original. Attribution of material, and permission to use it, is an obligation and necessity of everyone, historian, writer, or collector. If an item is "cut" or "edited" in a document it is important that this be identified and that the change described. The integrity of the original document must not be compromised.

Summary

Scanning gives us a wonderful tool for dealing with information and documents in ways that have never before be available. We can make copies that are better than the originals, and provide the product in formats that are more useful and available than ever before.

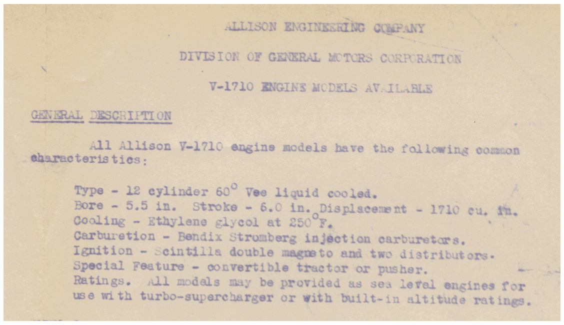
The equipment needed is readily available, and costs less than what it would cost to pay for one days Xeroxing at the National Archives. The associated computing power, memory, and storage media are likewise affordable while the software to allow full utilization of the capability is fun, though time consuming to use.

It is hoped that this article encourages the reader to develop scanning skills and thereby improve the access to rare and otherwise unavailable textual documents that are so important to capturing and sharing the history of aviation and aircraft engines.

Additional Reading

The Internet is a wonderful source of comments, techniques, evaluations and critiques on scanning and processing methods and equipment. The subject is evolving rapidly and you are encouraged to seek out specifics and experiences appropriate to the equipment and software you are using.

Daniel D. Whitney
Orangevale, CA 95662
January 30, 2002



ALLISON ENGINEERING COMPANY

DIVISION OF GENERAL MOTORS CORPORATION

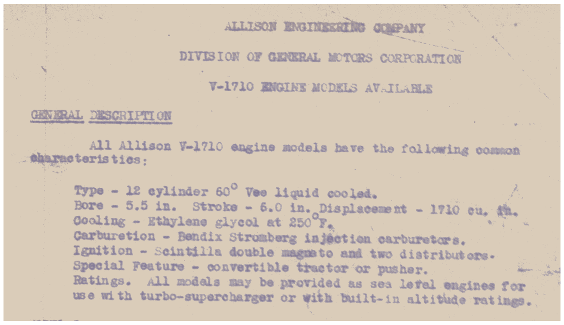
V-1710 ENGINE MODELS AVAILABLE

GENERAL DESCRIPTION

All Allison V-1710 engine models have the following common characteristics:

Type - 12 cylinder 60° Vee liquid cooled.
Bore - 5.5 in. Stroke - 6.0 in. Displacement - 1710 cu. in.
Cooling - Ethylene glycol at 250°F.
Carburetion - Bendix Stromberg injection carburetors.
Ignition - Scintilla double magneto and two distributors.
Special Feature - convertible tractor or pusher.
Ratings. All models may be provided as sea lefel engines for use with turbo-supercharger or with built-in altitude ratings.

Figure 1: Color Picture, 16 million colors, 4,729 KB
300 dpi, 1854x1052 pixels, 6.18"x3.5"



ALLISON ENGINEERING COMPANY

DIVISION OF GENERAL MOTORS CORPORATION

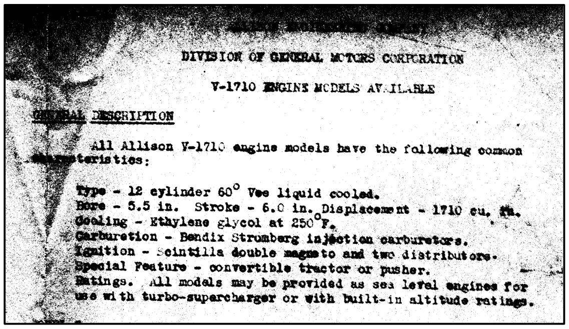
V-1710 ENGINE MODELS AVAILABLE

GENERAL DESCRIPTION

All Allison V-1710 engine models have the following common characteristics:

Type - 12 cylinder 60° Vee liquid cooled.
Bore - 5.5 in. Stroke - 6.0 in. Displacement - 1710 cu. in.
Cooling - Ethylene glycol at 250°F.
Carburetion - Bendix Stromberg injection carburetors.
Ignition - Scintilla double magneto and two distributors.
Special Feature - convertible tractor or pusher.
Ratings. All models may be provided as sea lefel engines for use with turbo-supercharger or with built-in altitude ratings.

Figure 2: Color Drawing, 16 colors, 109 KB
300 dpi, 1854x1052 pixels, 6.18"x3.5"



DIVISION OF GENERAL MOTORS CORPORATION

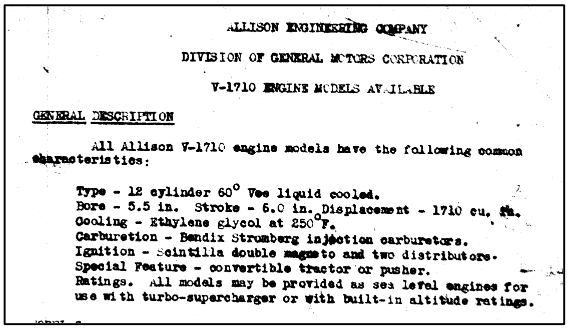
V-1710 ENGINE MODELS AVAILABLE

GENERAL DESCRIPTION

All Allison V-1710 engine models have the following common characteristics:

Type - 12 cylinder 60° Vee liquid cooled.
Bore - 5.5 in. Stroke - 6.0 in. Displacement - 1710 cu. in.
Cooling - Ethylene glycol at 250°F.
Carburetion - Bendix Stromberg injection carburetors.
Ignition - Scintilla double magneto and two distributors.
Special Feature - convertible tractor or pusher.
Ratings. All models may be provided as sea lefel engines for use with turbo-supercharger or with built-in altitude ratings.

Figure 3: B&W Drawing, 2 colors, 111 KB
300 dpi, 1854x1052 pixels, 6.18"x3.5"
Scanned using Blue channel



ALLISON ENGINEERING COMPANY

DIVISION OF GENERAL MOTORS CORPORATION

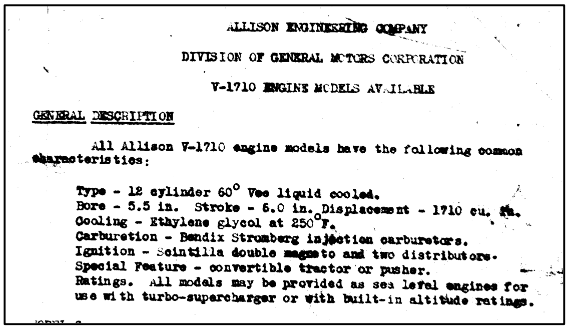
V-1710 ENGINE MODELS AVAILABLE

GENERAL DESCRIPTION

All Allison V-1710 engine models have the following common characteristics:

Type - 12 cylinder 60° Vee liquid cooled.
Bore - 5.5 in. Stroke - 6.0 in. Displacement - 1710 cu. in.
Cooling - Ethylene glycol at 250°F.
Carburetion - Bendix Stromberg injection carburetors.
Ignition - Scintilla double magneto and two distributors.
Special Feature - convertible tractor or pusher.
Ratings. All models may be provided as sea lefel engines for use with turbo-supercharger or with built-in altitude ratings.

Figure 4: B&W Drawing, 2 colors, 36 KB
300 dpi, 1854x1052 pixels, 6.18"x3.5"
Scanned using Green channel



ALLISON ENGINEERING COMPANY

DIVISION OF GENERAL MOTORS CORPORATION

V-1710 ENGINE MODELS AVAILABLE

GENERAL DESCRIPTION

All Allison V-1710 engine models have the following common characteristics:

Type - 12 cylinder 60° Vee liquid cooled.
Bore - 5.5 in. Stroke - 6.0 in. Displacement - 1710 cu. in.
Cooling - Ethylene glycol at 250°F.
Carburetion - Bendix Stromberg injection carburetors.
Ignition - Scintilla double magneto and two distributors.
Special Feature - convertible tractor or pusher.
Ratings. All models may be provided as sea lefel engines for use with turbo-supercharger or with built-in altitude ratings.

Figure 5: B&W Drawing, 2 colors, 35 KB
300 dpi, 1854x1052 pixels, 6.18"x3.5"
Scanned using Red channel

Alternative Raster File Formats, Benefits and Applications

	BMP	TIFF	GIF	JPEG
Scheme	Pixel x Pixel with pallet for up to 8 bit color, uses scan lines/rows for higher RGB values	Indexed or Tagged strips of horizontal and/or vertical data	Pallet based indexing with LZW compression, interlaced	Separates YCbCr Color and Luminance data by pixel, then 8x8 tiles to DCT algorithm,
Compression	RLE-Run Length Encoding	LZW-Dictionary based encoding of patterns in file with tile based storage	LZW using 8 bit pallet	Hi-freq data discarded and adjacent color/brightness data averaged. DCT algorithm operates on tiles. User sets quality
Compression Type	Lossless	Lossless	Lossy	Lossy
Compression Ratios	Not much	50% with LZW	Up to 40% lossless, lossy up to max of 5:1	10-20:1 with no visible loss in quality, up to 200:1 possible
Use	Good for Screen Shots	Good for print work. Good for line drawings.	Good for images with hard edges and sharp color changes	24 bit color, 8 bit gray. Not good for images with hard edges and sharp color changes.
Benefits	Fast and Efficient screen display	Works with RGB, LAB (YCbCr), CMYK	Good for web work needing <256 colors	Good for continuous tone photos, very little perceived quality loss
Limitations	Not good for continuous tone photos		Good for flat or limited colors, max 256 colors	Should "edit" using Lossless format. Not good for continuous tone or true colors
User Controls	None	PSP-Set DPI and LZW Compression	Set Transparency Level and version.	Set compression vs Quality at 1-100. Determines aggressiveness of DCT rounding of each tile, results is losslessly compressed using Huffman compression. Set level of compression.

Notes:

- LZQ is Lempel-Ziv-Welch, a substitutional or dictionary-based encoding algorithm that effectively creates its own look-up dictionary on the fly. Patterns of data are identified in the data stream and matched to entries in the dictionary, and if necessary a new code phrase added to the dictionary. RLE encoding compresses consecutively repeated values, whereas LZW compresses by identifying repeated combinations wherever they appear.
- RLE is Run-Length-Encoding, it works by taking repeated sequences and expressing them in two bytes so that a row of similar values is describes as the number of identical pixels and their color.
- DCT is Discrete Cosine Transformation, a mathematical algorithm that defines the equation representing the relevant brightness and color values in each 8x8 tile.
- LAB is name for the Luminance, Chrominance (YCbCr) scheme for describing the brightness and color of a given pixel, or group of pixels.
- A 1600x1200x16M-color photo of 5.5 MB, began as a 408 KB JPG, converted to: 5.6 MB BMP, 917 KB GIF, and 4.2 MB TIF.
- A 2640x4122x2-line drawing of 1.3 MB, began as a 272 KB TIF, converted to: 1.3 MB BMP, 262 KB GIF, and 2.9 MB JPG (with compression set at "2", and now has "10" colors, setting compression at "15" reduced to file to 1.6 MB, but now has "56" colors! Setting the dpi from 300 to 600 made no change in file sizes. "Colors" are artifacts of the JPG compression.).