DATA COMPRESSION TECHNIQUES

Brian Jordan*

Abstract: Data volumes relating to graphic images contained within an Electronic Pre-press System are high, and this leads to the requirement for large amounts of mass storage and extended processing and telecommunication times.

The following Paper quantifies the data volumes required when various classes of graphic-arts originals are digitised and stored within a Pre-press System and describes in outline two specific techniques for reducing those data volumes by an order of magnitude.

The advent of what can broadly be termed "Digital Image Processing Systems" or more precisely, "Electronic Pre-press Systems" for the Graphic Arts Industry has precipitated a careful review of the volumes of data which need to be stored and manipulated in those systems.

The handling of digitised type has been common place in the Graphic Arts Industry for some years. Typically, text processing systems code each character as one byte (8 bits) of information and one full page of text containing 20,000 characters can be represented in digital image storage by 20,000 bytes in addition to a relatively small overhead of composition commands, font calls etc.. Once we turn our attention to graphics however, the volumes of data required to represent an image rise dramatically. I have considered these data volumes under two main headings.

1. Pre-screened Graphics, Line-art and Scanned Text.

Pages containing a mixture of pre-screened graphics, line-art and text need to be scanned and digitised at a very high resolution, in order to maintain the fidelity of the smallest screened dots. For example, material containing 150 line per inch screened graphic information

* Technical Director, Crosfield Electronics Ltd.

will need to be scanned at a resolution of 1800 x 1800 pixels per inch in order to be sure of retaining the finest detail and highlight dots. As the scanned information has only two levels, i.e. black or white, only 1 bit per pixel is required to define that picture element.

One A4 page scanned at 1800 x 1800 picture elements per inch, with 4 colour separations would require approximately 142,000,000 bytes, i.e. a storage capacity of 142 Mbytes would be required to store one page.

2. Continuous Tone Graphics

If the origination material is in continuous tone form, then the amount of data required to store that image can be somewhat reduced by handling the data through the Pre-press System as continuous tone and only introducing the screen at the output device, e.g. on an Electronic Screening Colour Scanner.

If the final output is required to be at 150 line per inch screen ruling, then the imput scanning resolution will need to be 300 x 300 pixels per inch at same size enlargement. Pixels are digitised to 256 grey levels, which represents 1 byte of information per pixel per colour.

Therefore, one A4 page of colour graphics would require approximately 31,700,000 bytes, i.e. 31.7 Mbytes of storage capacity would be required for one A4 colour picture.

Problems Relating to Data Volumes

1. Data Storage

If text and graphics information is to be digitised and stored in an Electronic Pre-press System, then large amounts of digital mass storage devices and media are required. Figure 1 is a Table showing a comparison between the number of pages of pure text (stored in symbolic form), continuous tone graphics and line-work, which can be stored in standard 300 Mbyte disc packs and reels of 6250 bits per inch group encoded tape. As can be seen from the Table, while 300 Mbyte disc pack can store well over 1,000 pages of symbolically coded text, it can only store 8 A4 colour pages digitised from continuous tone originals and only two A4 colour pages digitised from pre-screened graphics, line-art or scanned text.

TYPE OF PAGE	APPROX. NUMBER OF PAGES STORED	
	300 Mbyte DISC PACK	6250 GE TAPE
TEXT	1250	625
CON-TONE	8	4
LINE	2	1

Fig. 1

2. Data Processing Times

Large data files exert heavy overheads on Image Processing facilities. One of the main overheads, relates to the time taken to physically read or write the data too or from the magnetic storage media. Reference to the Table in Figure 2, indicates the size of the problem. The read or write times specified in the Table, are the fastest possible times that could be achieved in reading data from disc or tape, assuming that the data was optimally organised on the magnetic media and that there was an infinitely large memory to absorb the data as it was being read.

Even under these ideal conditions, whereas a page of symbolically encoded text can be read in 30mS from disc, an A4 page of continuous tone colour will take over half a minute to read and a page of high resolution line-work will take more than two and a half minutes to read. When it is taken into account that data can never be manipulated in blocks even approaching several tens of megabytes, it can be seen that the input output overhead in storing and re-trieving data can soon become a very high percentage of any image processing time.

TYPE OF PAGE	TIME TAKEN TO READ (WRITE) 1 PAGE	
	300 Mbyte DISC	6250 GE TAPE
TEXT	.03 secs	0.036 secs
CON-TONE	35 secs	43 secs
LINE	156 secs	192 secs

Fig. 2

3. Communication Systems

There is an increasing trend amongst the larger Printers and Publishers to offset the time and transportation costs involved in shipping large quantities of printed matter around the world by setting up remote printing sites and transmitting composed page information from a central site directly to the remote printing sites, using satellite or other appropriate communication links.

As can be seen from the Table in Figure 3, only wide band and consequently expensive communications links provide transmission times, which in anyway would be acceptable for the majority of publishing work again with the exception of symbolically encoded text.

TYPE OF PAGE	TRANSMISSION TIMES	
	56 kbit line	1.533 Mbit link
TEXT	3 secs	0.1 secs
CON-TONE	75 mins	2.7 mins
LINE	338 mins	12.4 mins

Fig. 3

Data Compression

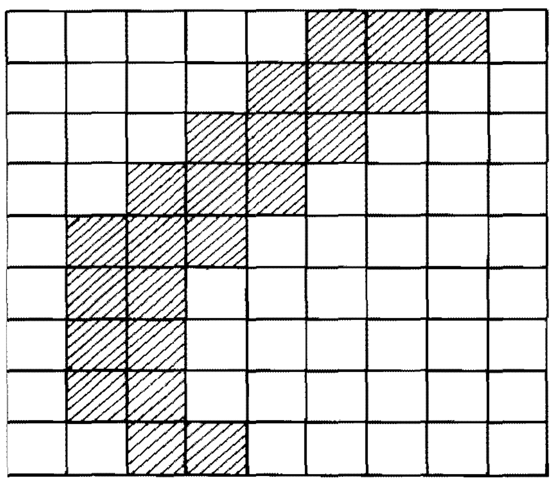
1. The Objective

It is becoming possible through the application of relatively complex Data Compression algorithms and the availability of very high speed processing elements and low cost memory to design hardware capable of running at disc transfer rates which will enable data volumes to be reduced by an order of magnitude.

The rest of this Paper describes in outline two approaches being implemented by Crosfield Electronics, one relating to the compression of line-work and the second, relating to the compression of continuous tone coloured images.

2. Data Compression of "Line-work"

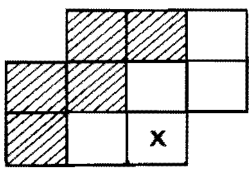
When line-work is scanned, as previously described, a pixel map is built-up (see Figure 4 at high resolution) (1800 pixels x 1800 pixels) on a scan-line by scan-line basis. Simple encoding techniques could be applied to such data, for example, simple run length encoding whereby, instead of the individual pixels being stored, the length of the runs of black or white pixels are recorded. However as the run lengths of black or white pixels become very short, as would be the case with half tone graphics of fine text, then the "compression" that can be achieved approaches unity or can even result in more data being recorded rather than less.



PIXEL MAP

Fig. 4

A more complex algorithm is therefore required, which basically attempts to predict the next pixel in a sequence of pixels, taking into account the value of the surrounding pixels. During de-compression and re-constitution of the original image, the same prediction algorithm is applied, hence, data only needs to be stored when a wrong prediction is made. Figure 5 shows a typical prediction window; as the data isscanned a few lines are stored and then the data is examined through the prediction window, which moves progressively down the scan lines. The predicted value of the pixel marked 'X' which is made by the prediction algorithm, is compared with the actual value of 'X' and data is only stored if the prediction is wrong. In the example shown in Figure 5, there would be a high probability that the value of pixel 'X' would be white. A number of prediction algorithms are used simultaneously, these algorithms being tuned to provide the best predictions on various types of data, e.g. screen line or text. The predictor producing the most accurate results for a given block of data is also recorded with that block, so that upon de-compression, the appropriate predictor can be used to re-constitute the data.

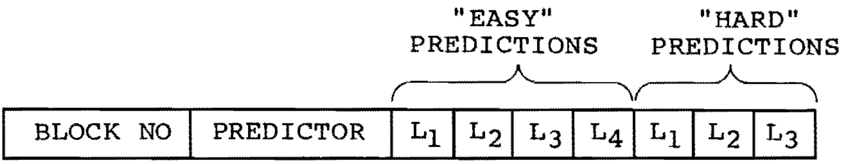


PREDICTION WINDOW

Fig. 5

In addition to predicting the value of the next pixel, the predictor also determines whether the prediction was "easy" or "hard". It records and "easy" verdict if the probability of a correct predictor is high and a "hard" verdict if the probability of the correct prediction is low.

Compressed Data is therefore recorded as shown in Figure 6 being the block number code depicting the predictor that is used, the run lengths of the "easy" predictions and the run lengths of the "hard" predictions. Run lengths are coded in accordance with the standard Huffman code. The advantage of grouping the "easy" predictions together, is that, relatively long run lengths are generated enhancing the degree of compression obtained.



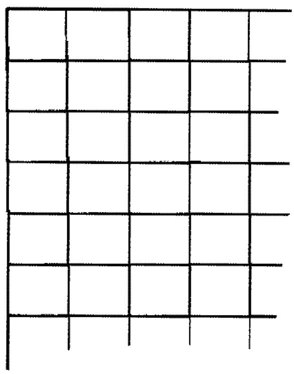
COMPRESSED DATA BLOCK FORMAT

Fig. 6

Using these techniques, compression ratios in the order of 6 or 7:1 are typical for half tone areas, while in areas of text and line-art, compression ratios in the 20 or 30:1 range are achievable. Although it is always difficult to define an "average" page, the overall compression ratio achieved by these technqiues, is in the order of 13:1 which is in line with our overall objective of an order of magnitude reduction in data volumes.

3. Continuous Tone Compression

As it can be seen from the typical data volumes shown for an A4 page, earlier in this Paper, graphic material can be more efficiently stored in continuous tone form than in line form. When the continuous tone data is scanned, a pixel map is still obtained, but this time each picture element at a lower spacial resolution (typically 300 x 300 picture elements per inch) is represented by 8 bits of information per colour, representing 256 grey levels per colour



PIXEL MAP
(8 bits/PIXEL/COLOUR)

Fig. 7

Whilst the de-compression and re-constitution of the image data using the technique previously described for line-work will result in an exact replica of the data before it was compressed, the approach taken for continuous tone compression is effectively to "throw away" data, in order to achieve compression, but to ensure that when the data is re-constituted, the ffect of the data loss is not noticeable to the eye when the data is output to a film or printing forme. An outline of the compression technique is described below.

Suitable sized blocks of the original image data are transformed through a complex algorithm from discrete pixel grey level values into a map of the frequency components of the image, starting with the D.C. component at the top left hand corner of the block, with higher frequency values recorded as one moves away from the D.C. component in the horizontal or vertical axis. (See Figure 8). Up to this stage, there has been no data reduction, but the data has been transformed to a format which now allows for classification and data reduction. The transformed data is now classified into a number of classes depending on the data content and in accordance with the class, the data can be thresholded and re-quantised. Finally, the data is recorded in compressed form as the D.C. component with the higher frequency components coded using a Huffman code.

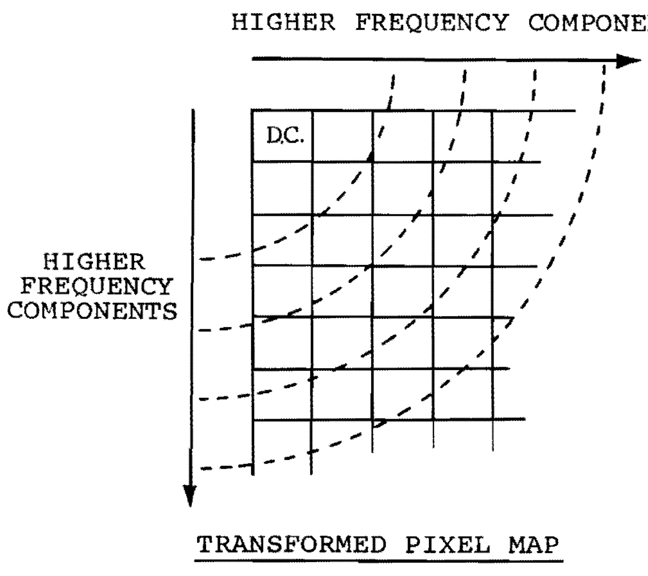


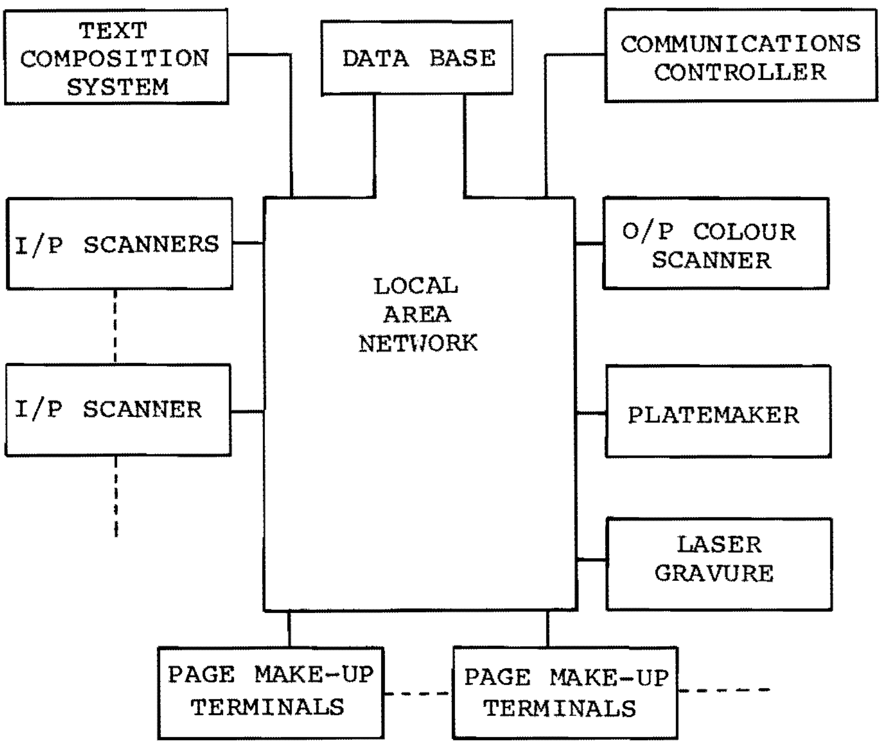
Fig. 8

Using these techniques, compression ratios in the 8 or 10:1 area can be achieved without detectable quality degradation.

Advantages of Data Compression

From the discussion earlier in this Paper, regarding data volumes, storage and processing requirements and telecommunication timings, it is obvious that compression of an order of magnitude will make a dramatic reduction in all areas.

The ultimate objective is shown in Figure 9. This represents a fully "networked" Electronic Pre-press System, where the individual peripherals of the Pre-press System, input and output scanners, text front end systems, page make-up terminals, communications controllers and printing form output devices such as laser-platemakers and laser-gravure cylinder mating equipments can all be sited on some form of local areas network and communicate with one and other through a common database of text and graphics. While it is possible to configure such a system on current technology, the online data storage requirements for the database and the overheads in transmitting the high data volumes around the local area network, make large systems of this type currently impractical. The introduction of Data Compression or De-compression at each of the peripherals and within the database, will make such a system a practical reality.



NETWORKED PRE-PRESS SYSTEM

Fig. 9