© von Karman Institute for Fluid Dynamics
Waterloosteenweg 72 – 1640 Sint-Genesisu-Rode

Dear reader,

The electronic document attached is protected by copyright laws.

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

Please do not distribute the document to third parties outside your workgroup as under current copyright law it is only intended for your personal scientific use. If you do not comply with the above conditions of use, you may be held liable for copyright infringement.

VON KARMAN INSTITUTE FOR FLUID DYNAMICS

LECTURE SERIES 1984-03

DIGITAL IMAGE PROCESSING IN FLUID DYNAMICS

MARCH 5-9, 1984

PLANNING AN EXPERIMENT FOR DIGITAL IMAGE PROCESSING

JAVIER JIMENEZ - IBM SCIENTIFIC CENTRE, MADRID, SPAIN

PLANNING AN EXPERIMENT FOR DIGITAL IMAGE PROCESSING

Javier Jimenez
IBM Scientific Centre
P. Castellana 4, Madrid (1), Spain

When a digital image processing (DIP) system is used in an experiment, it should be seen as a (fairly expensive) measuring instrument, and planned accordingly. Its main components are outlined in figure 1. In these notes we will try to determine some of the design criteria for the different parts of the system as a function of their intended use, and to give a short summary of the options which are available at present for each of them. The level of technology that we try to cover is roughly what can be found commercially or built from commercially available components. Custom designs and experimental technologies will not be treated in most cases.

The core of the system is, of course, the experiment itself. Its design depends on the phenomenon being studied and is not addressed here. We will assume that, if DIP is going to be of any use, some of the data in the experiment are acquired in the form of flow visualisation. A careful study should be made on whether the visualisation data really require digital processing. Photography is a well developed technique whose capacity to acquire and store data is still unmatched by digital means. No sensible digital system can, at present, acquire data at a speed comparable to a standard 16 mm cine camera for any length of time, let alone any of the available high speed photographic systems. Moreover, the photographic emulsion is still the long term storage medium with a highest density. An economic analysis is, of course, even more in favour of analog photography.

The rationale for digital processing is usually the need for quantitative analysis of the images, or the use of specialised sensors with capabilities that go beyond normal photography. Even in these cases the possibility should be

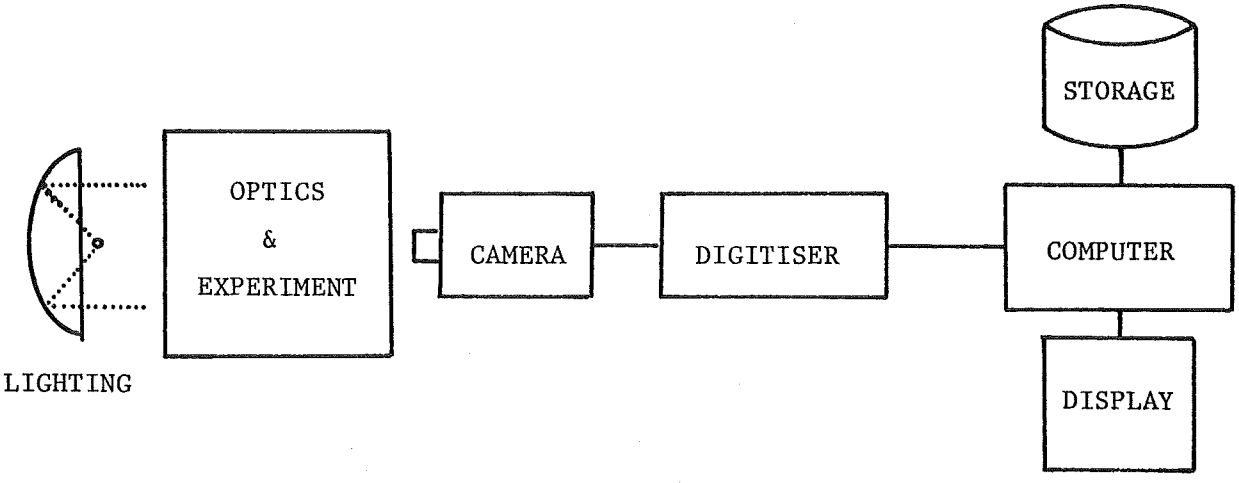


Fig.1- Principal components to be considered in designing a laboratory digital image processing system.

considered of using photography for data acquisition, followed by digitisation and analysis of selected photographic negatives. This procedure shares many of the advantages of analog acquisition with those of DIP, but it introduces an extra step in the data path, with associated extra calibration procedures and errors.

COMPUTER REQUIREMENTS

The second most important item in a DIP system is the computer. The overriding characteristic of a digital image is its size. A 35 mm photographic negative with a moderate resolving power of 50 line pairs/mm is approximately equivalent to a 2400x3600 image with 5 bits/pixel, containing nearly nine million pixels and requiring 5.4 Mbytes of storage. Such an image is probably too big for most applications. The size of an image, and therefore the amount of computer power needed to process it, is a function of its format (linear dimension in pixels, if we assume a square image), its depth (bits needed to code each pixel) and the number of operations per pixel needed for the desired process. Format and depth define the amount of storage; format and operations/pixel define the total number of operations needed in the processing, and they define the speed of the processor required.

Figure 2 is a plot of format versus depth. The areas marked define sizes that are normally associated with different applications of DIP, and the plot contains isolines of constant total storage. Two of them are especially important; most small computers are limited to 16 bit addressing, and the largest amount of storage that can be treated as a single array without taking special precautions is 64 Kbytes. This is also the largest image that can be managed "globally" in those systems without complicating the software. This does not mean that "local" operations, such as contrast manipulation and filtering, cannot be done on larger images, but they require working from secondary storage, which is much slower. Global operations, such as contour following and pattern recognition, become complicated unless the complete image can be held in a single memory page. The other important line is 4 Mbyte. Although it does not represent any fundamental restriction, it is typical of the maximum amount of real memory present in laboratory "superminis". Most of these machines have virtual memory management systems that are transparent to the user and address up to 16 Mbytes, but performance falls with increasing workspace size. The region above 4 Mbytes is at present the realm of large mainframes and special purpose machines.

Figure 3 is a similar plot of size versus operations/pixel. On the right hand side, the vertical axis is marked with a few selected algorithms. The number of operations given here for those algorithms represents the inner loop of a careful assembler implementation; complete programs would probably have higher operations counts, especially if they are written using a high level language. The different shaded areas

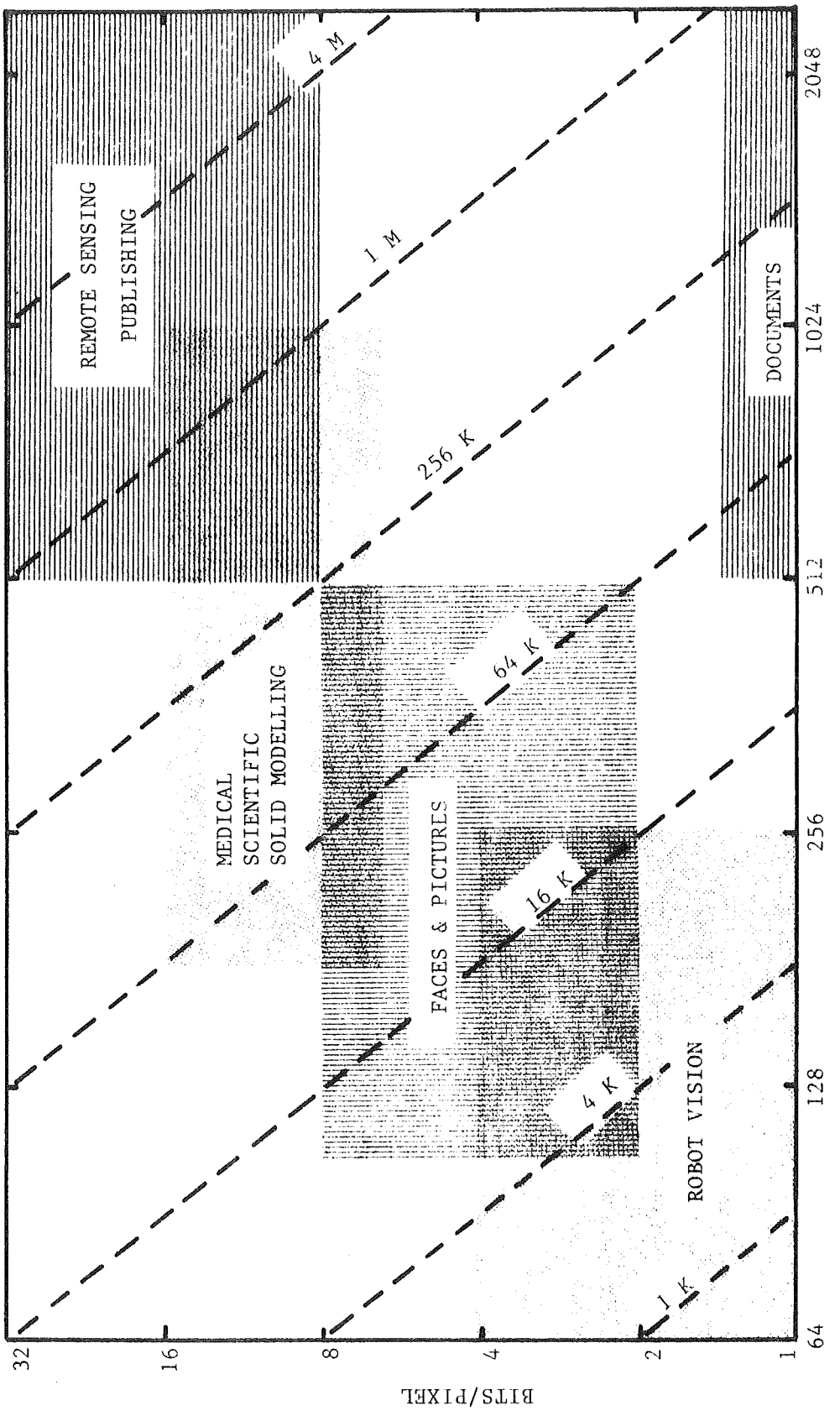


Figure 2.- Classification of different DIP applications in terms of image format and pixel depth. Dashed lines represent total storage needed for the image, in 8 bit bytes.

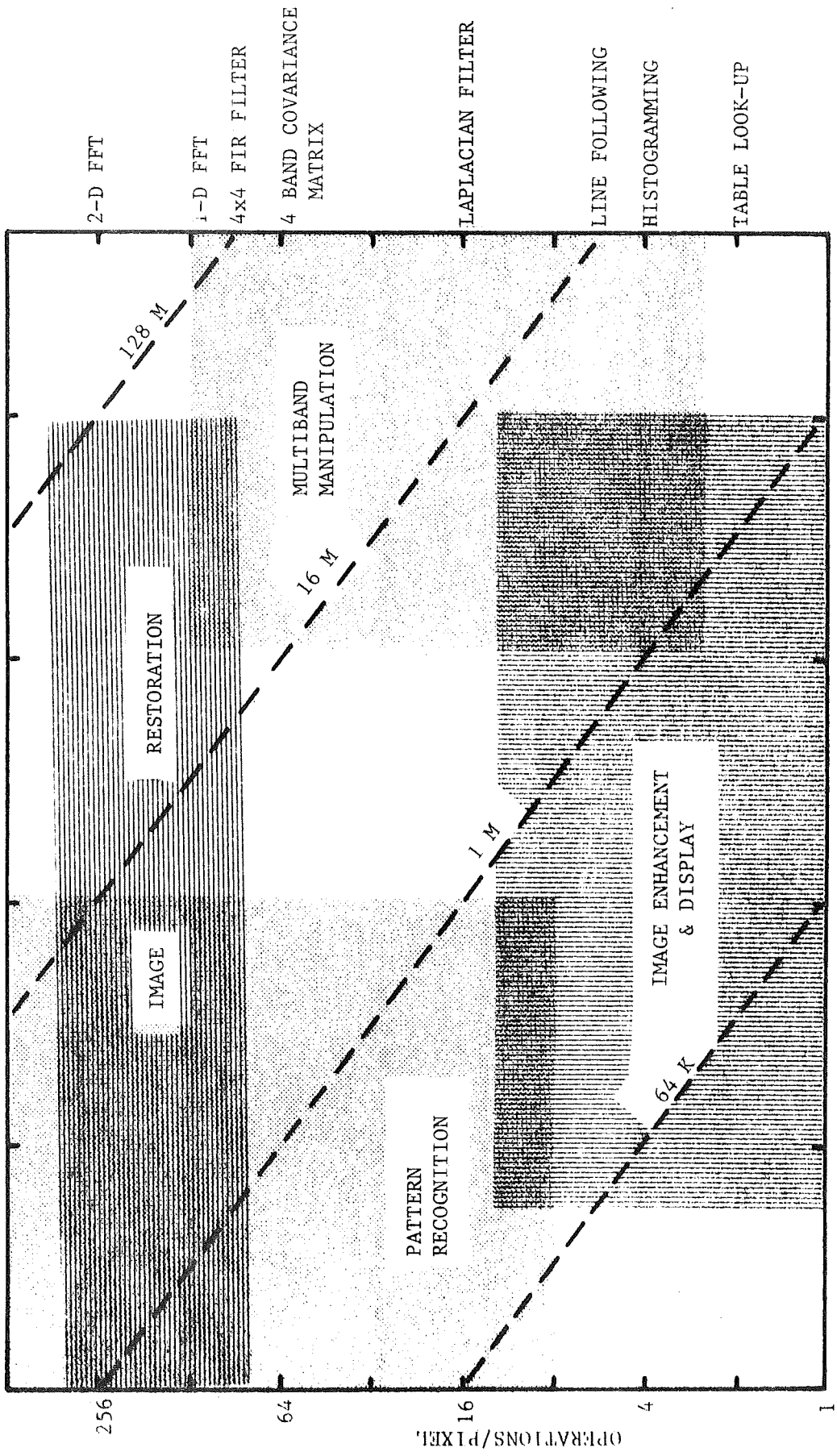


Figure 3.- Classification of different DIP activities in terms of image format and number of operation needed for each pixel. Dashed lines represent total operations needed to process one image.

those algorithms represents the inner loop of a careful assembler implementation; complete programs would probably have higher operations counts, especially if they are written using a high level language. The different shaded areas represent here different processing activities and the isolines are total operations needed to process a whole image. These isolines should be read in terms of processing time. The microprocessors found in personal computers and minis are usually rated near 0.2-0.8 MIPS (million instructions per second), superminis and mainframes near 1-2 MIPS, and special purpose machines can go up to 10-20 MIPS. Thus, a process using 10 million operations would take 5-10 sec in a mainframe, but up a to minute in a minicomputer and less than a second in a special purpose machine. The size of the images used in an experiment should be chosen so that they can be processed in a reasonable time with the available equipment. In particular note that it is seldom possible to do any real time digital image processing using anything but special purpose machines.

Another area that needs careful consideration is the interface of the input digitiser to the computer. Communication lines are usually slower than CPU's, and most of the time needed to process an image can be spent moving it in and out of central storage. The optimum solution is to have a dedicated DMA port to the system bus. Transfer speeds are then close to the speed of the processor and software overhead is low. In some systems it might not be possible to dedicate a whole DMA channel to a single instrument; most digitisers can be configured to connect to any of the standard multi-instrument bus interfaces, of which the fastest is probably the IEEE488 (1). This is basically an external bus connected to an interface port on which several instruments (up to 15) can transmit or receive information. Its maximum rated transfer speed is 1 Mbyte/sec, according to the standard, but depends on cable length (the standard maximum is 20 m). Software overhead depends on the number and types of instruments in the bus. Simple block transfers are rather fast, but switching among instruments and between talker and listener modes is complicated and can add a substantial overhead to the transmission time per pixel. Moreover, the actual transfer speed for of a given IEEE488 controller depends on the manufacturer and may be considerably slower than the standard.

Most instrument interfaces are only available for reasonably small minicomputers which can only hold a small number of complete images in memory. In those cases the limiting transmission speed is that of transferring the images to mass storage or uploading them to a larger computer. Disk transfer can seldom be done at more than 100 Kbytes/sec sustained speed and communications between minis and mainframes is usually limited by the multi-user environment of the mainframe. In practice 100 Kbytes/sec should be considered an absolute maximum input rate for a sequence of images. Thus, a 500x500 image at 8 bits/pixel needs at least

2.5 sec. Since DMA's usually address whole bytes, decreasing the number of bits per pixel does not necessarily increase the input rate unless special hardware is available to do the packing of several pixels into a single byte.

Another limiting factor is the amount of secondary storage available. The speed estimates in the last paragraph assumed at least a Winchester type hard disk. However most disks available for minicomputers are limited to a few tens of Mbytes and would only hold short sequences of average size images. Tapes are slower, and more expensive, than disks, but they hold more data. Diskettes are too slow and do not hold enough data for any but the most trivial DIP. As with processing power, an experiment should be planned taking into account the available mass storage. This planning should also consider the problems of long term archiving and back-up.

CAMERA AND DIGITISER.

Before an image can be digitally processed it has to be acquired by a sensor and digitised. Acquisition systems take many forms, ranging from simple manual tablets to CCD cameras and scanning microdensitometers. Their price and performance vary widely and their choice for a particular experiment is an important aspect of system design. Image acquisition systems can be classified broadly as "real time" or "static", depending on the working speed.

Real time systems are essentially video cameras and they are intended to sense, digitise and transmit an image in a time of the order of the characteristic experimental time scale. The classification depends, therefore, on the experiment being planned but, in practice, real time means approximately the commercial video refresh rate of 25-30 frames/sec. Higher rates are difficult to achieve and much lower ones are usually too low for fluid mechanics. The choice of a commercially compatible refresh rate, and format, has the additional advantage that the image can be easily monitored on a standard TV screen while it is being acquired. This is especially useful in such, otherwise "messy", activities as framing, focusing, etc.

It is often found, however, that video rate cameras have performance characteristics which are insufficient for quantitative work. The alternative is then to acquire the images in real time by other means, usually photographic, and convert them to digital form with a slow system. There are instrumentation cameras that can digitise an image with good accuracy in times ranging from a few seconds to a few minutes. Scanning microdensitometers are slower (5 min to hours), but they are more precise; they are also much more expensive. Digitising tablets are the most economic option and in some cases the most convenient, but they are slow, they have to be operated manually and are only able to record linear features and yes/no information. A summary of typical characteristics

of the different systems is given in Tables 1 and 2. We will discuss them now in some detail.

Tablets

They take the form of a drawing or copying table. A cursor with a loupe, similar to the "mouse" in some personal computers, allows the operator to select a point and transmit its coordinates to the computer by pressing a button. There is usually a "tracking" mode that transmits the coordinates as soon as the cursor moves by a predetermined amount, allowing for continuous tracing of lines. A picture is fixed to the table, registered by taking the coordinates of at least three points, and the coordinates of important lines, points, features, etc. are fed into the computer. The system has limited capacity to code for colour or intensity and cannot be used to digitise extended regions, but it is useful in applications like the digitisation of particle tracks, outlines of regions, etc.

The main advantage of tablets is their relatively low cost, and the fact that the operator can exercise his judgement in choosing important features. This last point can also be a disadvantage since it introduces a subjective component in the measurement. Tablets can be found in sizes up to 1x1 m, and the system has been adapted to film readers by means of more elaborate optical arrangements. The detection of coordinates is done by a variety of means that usually involve some sort of grid buried into the tablet. Digitisation speeds are of the order of a few points per second, but are mostly limited by the human operator. Connection to the computer is usually through a standard asynchronous RS232C serial interface. The rated accuracy of a good quality tablet is about 0.1 mm, but the reliability of human operators to locate a point is found experimentally to be only 0.2-0.3 mm.

Microdensitometers

Scanning microdensitometers are the highest precision instruments available to convert an image, usually in the form of a photographic negative, to digital form. A flat-bed scanner takes the form of a moving table on which the negative is mounted. A light beam is first focused on a calibrated entry window, then on the negative, and finally on another calibrated exit window. The light transmitted by this arrangement is measured by a photomultiplier tube, digitised and stored (see figure 4). The table moves under control of a small computer and scans different parts of the negative under the stationary optical system. The fact that the optics do not move during the scan produces superb characteristics of geometric accuracy and measurement stability.

Drum scanners work on the same principle, but the photograph is mounted on a rotating drum. The light source and detector are mounted on a lead screw along the drum and scan a spiral path on the subject. Because the sensor assembly has to

TABLE 1

Some typical characteristics of image sensors.

Sensor	Max. Dimension (mm)	Resolution (pix)	Maximum geometric distortion	Price ($)	Time per frame
Digit. Tablet	1000x1000	5000x5000	.2 mm	1,000 to 10,000	5 min to hours
Scanner Drum	1000x1000	100,000 x 100,000	0.01 mm	25,000 to 200,000	1-30 min
Scanner Flat-bed			0.001 mm		5-300 min
35 mm photography	24x36	2000x3000	1 %	1,000	variable
Vidicon: Sb_2S_3 Chaln.	10x10	900x900	1 %	3,000	1/30 sec to 1 min
		800x800		5,000	
Pb O		700x700		8,000	
SIT		500x500	3 %	10,000	
Matrix CCD	6x8	400x400	0.1 pixel	10,000	
Scanning CCD	30x30	2000x3000	1 pixel	7,000	2-100 sec

TABLE 2

Typical electrical characteristics of image sensors.

Sensor	Bits pixel	gamma	Lag	S/N (db)	Photometric uniformity (%)	Dynamic illumination range
Scanner: Drum	6	1	0	50	0.1	1,000
Scanner: Flat-bed	8	1	0	65	0.01	10,000
35 mm photography	5	0.6-1.2	0	45	20	100
Vidicon: Sb_2S_3	4	0.65	20	40	15	100
Vidicon: Chaln.		0.95	20		10	100
Vidicon: Pb O		0.95	3		10	200
SIT	3	1	10	30	20	50
Matrix CCD	7	1	0	60	10	100 (at 25°C)
Scanning CCD						5000 (at 0°C)

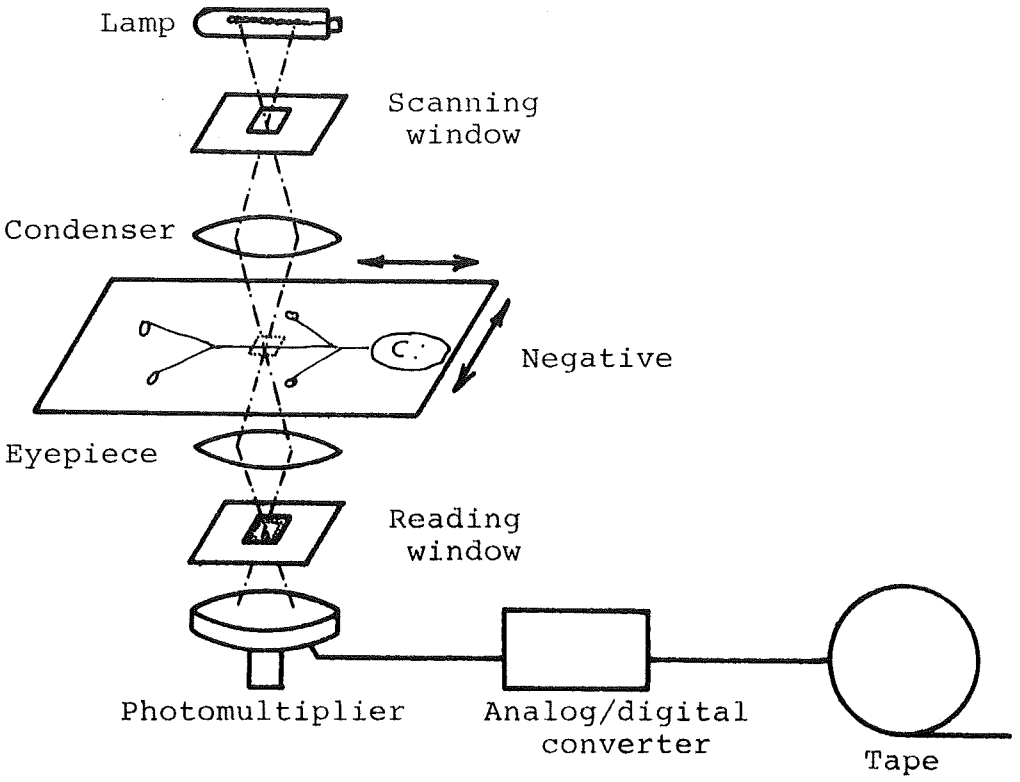


Figure 4.- Functional scheme of a flat-bed scanning microdensitometer.

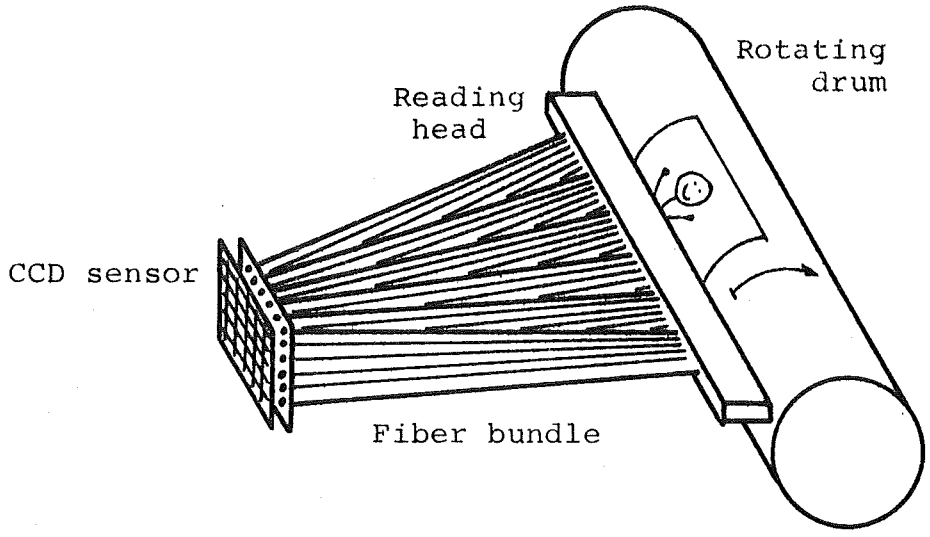


Figure 5.- Drum scanner showing the construction of an optical fiber head

move, the positional and photometric accuracy is lower than on flat-bed models, but drum scanners are usually faster and, generally, somewhat cheaper.

A relatively recent development is the optical fibre head. A linear array of up to 10,000 optical fibres, spaced 0.1 mm apart, is positioned along the drum of a scanner. The fibres are then rearranged into a 100x100 bundle that is presented to a solid state camera (figure 5). Thus, the illumination along a complete image row can be sensed simultaneously with good geometric accuracy. Because light absorption in the fibres is substantial and tends to be different among fibres, the photometric accuracy of these systems is low, and they are normally used as high speed, high resolution, one bit deep readers of industrial blueprints and documents.

Cameras

We include under this heading a series of devices with a mechanical/optical packaging similar to a photographic camera, but using a variety of light sensors. They include vidicons, CCD arrays and photographic film. We will not treat film directly in these notes; photography is a mature technology and films can be found in a wide range of types. The entries corresponding to photography in tables 1 and 2 are intended for use as a rough comparison with other sensors and refer to general use commercial films such as Kodak TRI-X. Good reviews of the general characteristics of film and photographic systems are found in (2,3). We discuss here the general response of camera sensors to light. The discussion is centred on vidicon targets but similar considerations apply to film or photodiode arrays.

Cameras are usually integrating devices, and they respond to the total light energy that is absorbed by their sensors during exposure. A typical response curve to static illumination is shown in figure 6. The horizontal axis is the logarithm of total absorbed energy, and the vertical one is the logarithm of signal intensity. At low light levels, the signal is dominated by the "dark current" which, in a vidicon, is produced by electron leakage in the target in the absence of illumination. This dark current increases with increasing temperature, usually doubling every 5-10 °C, and effectively limits the lowest light intensity that can be detected with a sensor. It is usually difficult to cool vidicon targets, which are only accessible from the front where the cooling equipment would block the incoming light, but silicon arrays, which have a relatively high dark current, can, and often are, cooled, for low light level work. Dark signal in a film is referred to as "fog" and is influenced by many factors besides temperature.

Above the dark current level, the relationship between the logarithms of input and output is approximately linear. The slope of the line is called the "gamma". A sensor with a high gamma gives a nonlinear response in which high intensities are

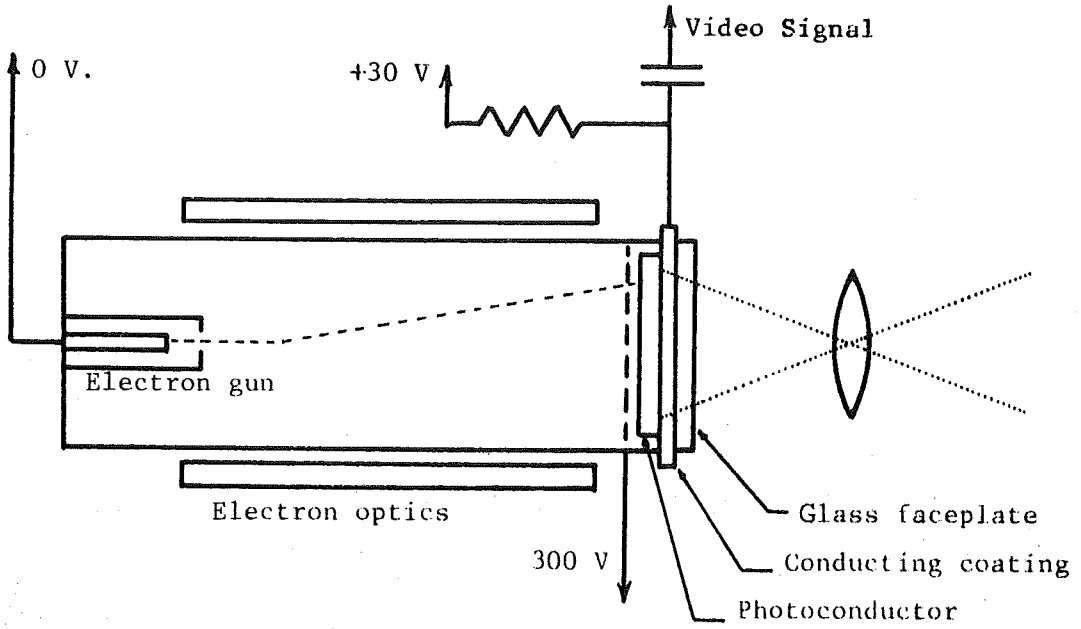


Figure 7.- Operating principle of a vidicon camera

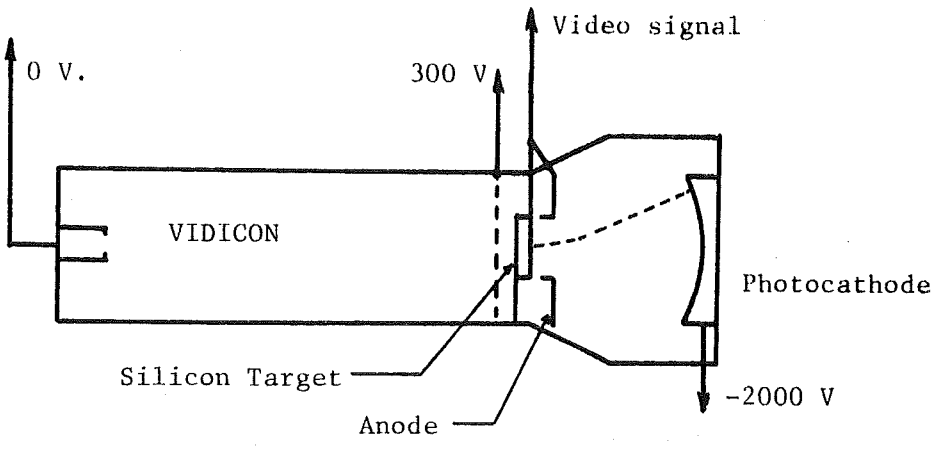


Figure 8.- General construction of a Silicon Intensified Target (SIT).

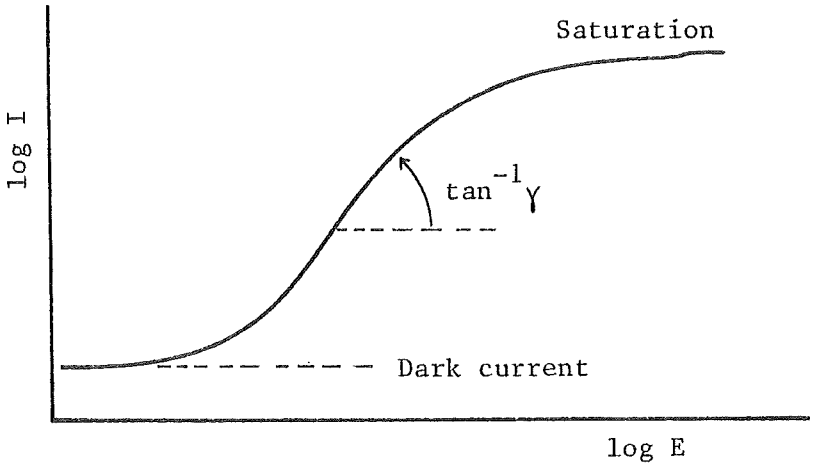


Fig.6- Typical sensor response to incoming light. Signal intensity refers to output current in photoelectric sensors, and to optical transmissivity in photographic film.

privileged while low intensities are difficult to distinguish. Low gammas produce the opposite result. A gamma equal to one represents a linear response of the sensor and is usually preferred in quantitative work.

Above a certain light level the sensor saturates and the output remains constant. This regime is dangerous for most electro-optical instrumentation. Since the output current has to flow through a relatively insulating photoconductor, high currents produce "burns" that leave permanent insensitive marks. This is true as much of most vidicons as of the cathodes of the photomultipliers which are the sensing elements in most scanners. One of the main advantages of solid state sensors is their relative resistance to burning.

Vidicons

These are cameras in which a photosensitive surface is read by an electron beam inside an image tube (4). The basic arrangement is shown in figure 7. A photoconducting storage target is deposited on the front face of the tube, covered on the front by a transparent conducting electrode and by a glass sealing plate, and exposed in the back to a scanning electron beam. In darkness the target is a good insulator and the electron beam charges it to cathode voltage. When light falls on the target it becomes conducting and leaks electrons to the front electrode, thereby accumulating a local positive charge which is proportional to the integrated light intensity between readouts. When a point of the target is scanned by the electron beam, it is recharged. The current thus generated is the output signal. Because, at normal scanning speeds, the beam might not be able to recharge completely the target, vidicons tend to keep some memory of the image between

successive scans, leaving an after-image that lags behind bright moving objects and that can be very bothersome in quantitative image acquisition. "Lag" is usually measured as percent signal remaining after the third scan and is given in table 2 for different target composition.

The characteristics of vidicons depend on the material of their targets. Four typical targets are represented in tables 1 and 2. Sb_2S_3 is the most common, and the cheapest, target. It is significantly nonlinear, but it is widely used in commercial applications because its low gamma matches well the high gamma (approx. 1.45) of the phosphors in commercial TV screens. It is an easy material to manufacture and SbS tubes have good resolution and geometric characteristics. However, its photometric characteristics are poor, it drifts, has a high lag and should be avoided for most quantitative work.

Chalnicon is a commercial name for a vidicon using targets from a family of compounds and mixtures of selenium, tellurium, arsenic and other related elements. They have high sensitivity, gamma near unity, are relatively immune to burns and have low dark current and good stability. They are also relatively inexpensive and a good compromise for image acquisition. Another material is PbO (Plumbicons). These targets are doped to form p-n junctions that work in photovoltaic rather than photoconducting mode. As a consequence they have very low lag and low dark current. PbO targets show some lateral light leakage and their resolution is relatively poor. They are also more expensive than the previous types.

An interesting type of target that merits some discussion is the Silicon Intensified Target (SIT). In these tubes, the faceplate is covered with a photocathode which emits electron as a response to light. This electrons are accelerated by a strong electric field and allowed to hit a silicon photoconductor (figure 8). The rest works like a normal vidicon. Because energy is given to the electrons before they hit the target, their effectiveness is greatly increased, and these tubes have sensitivities which are much higher (200+) than normal vidicons. Their geometric quality is relatively poor because it is difficult to provide a good focusing mechanism for the photocathode electrons. There are many more target materials and constructions that cannot be treated here. The best references are found in manufacturer specifications.

Photodiode arrays.

A photodiode works by accumulating electrons from electron hole pairs produced by incoming light. Therefore it is an integrating sensor in the same way that a vidicon target. It is possible to build arrays of silicon photodiodes in which each site corresponds to an individual pixel of the image. These arrays have several advantages ranging from the geometric stability inherent in their construction in a solid

matrix, to their linear response and good signal to noise behaviour. They have a relatively high dark current, but they are small devices which can easily be cooled from the back. The problem of how to scan them was solved with the construction of CCD analog registers.

A Charge Coupled Device (CCD) register is a linear array of capacitors implemented in silicon. The capacitance of the individual capacitors can be varied by changing the electronic properties of each site by means of individual voltage control lines. Each memory cell in the register consist of at least two capacitors. By controlling appropriately the capacitance of alternating sites in the row each capacitors transfers its charge to its neighbour. In a complete cycle, the charge contained in each cell shifts one position along the register; the charge contained in the last cell of the array is transferred to an output line (figure 9). In this way the complete register can be read sequentially. A linear CCD photosensitive array consists of an array of photodiodes, a series of transfer switches connecting each diode to a cell in a CCD register and lines for shifting and reading the register (see figure 10). The whole device is integrated in a single silicon chip. During exposure the transfer switches are closed and the diodes accumulate charge in proportion to the total integrated light energy that they receive; then, during a short read interval, the switches open and the charge in each diode is transferred to its associated cell in the register. Once the switches close the diodes are isolated from the register and exposure starts again. During exposure time, the CCD register is read into the computer.

The limiting factor in the accuracy of the system is the efficiency of charge transfer from one cell to the next. Assume that, in a single transfer, the charge transferred to the neighbouring cell is only 99% of the original; the charge in the last cell of a linear 1024 array has to go through the whole register before being read and its final transfer efficiency would only be 0.003%. Transfer efficiency decreases with reading speed, but modern implementations achieve acceptable reading accuracies with transfer rates in the order of megahertzs. A related problem is the generation of thermal electrons in the CCD cells. These electrons are added cumulatively to the transferred charge and are responsible for the relatively high dark current of these devices.

Linear arrays are used in image acquisition either directly when the subject moves under the array, as in drum scanners or in flowing systems, or in the form of mechanically scanning cameras. These cameras have a linear array that moves perpendicularly to its length in the focal plane of the lens. In this way one image axis is scanned electronically and the other mechanically. Since linear array with 2048 diodes are easily available, these cameras can produce images in the range of 2000x3000 which are scanned in about 2 sec. Their geometric accuracy in the axis of the array is very good, but the tolerance on the other axis is usually lower due to

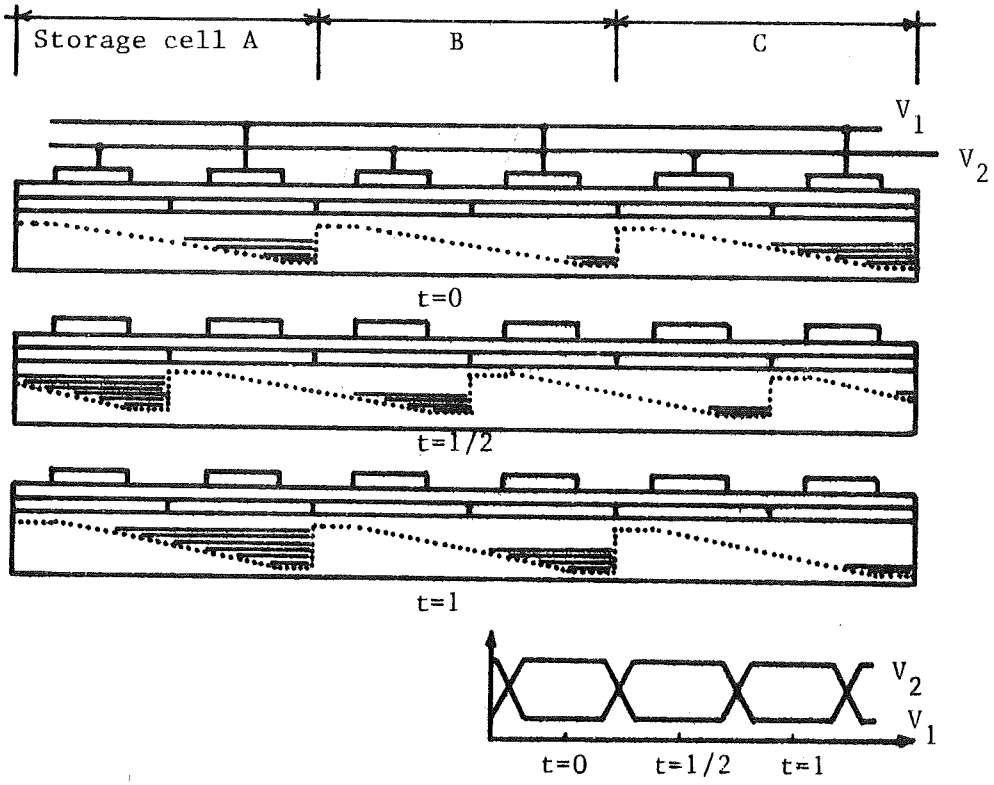


Figure 9.- A CCD register stores electric charge in potential wells controlled by surface electrodes. The charges are shifted among cells by changing the potential.

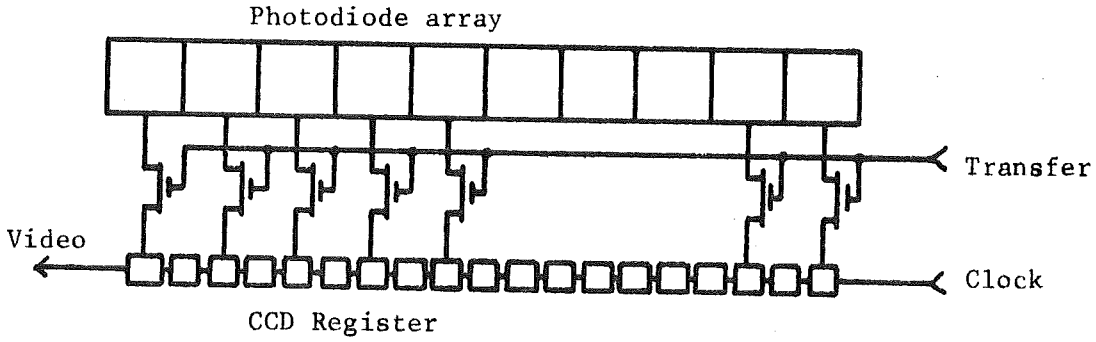


Figure 10.- A photodiode linear array. Diodes are the light sensing elements. The CCD register acts as a reading buffer.

mechanical play in the driving mechanism. Note that the transfer velocity implied by this scanning speed is 3 Mbytes/sec, which is usually too high for data acquisition systems.

It is also possible to find two dimensional CCD arrays which are scanned electronically in both axis. Their operating principle and signal characteristics are the same as in linear arrays, but their geometric accuracy is excellent on both axis. This technology is still state of the art and the largest cameras available are in the 300x400 range, chosen to be compatible with commercial TV. These cameras are still relatively expensive but they will probably develop into the choice image acquisition technology in the immediate future.

The main operational problem of CCD arrays is the lack of uniform sensitivity among the diodes. Arrays, especially the two dimensional ones, are offered in different quality grades depending on the maximum number of acceptable point defects (insensitive elements) and the size of the largest insensitive cluster; the price increases strongly with quality. A point is considered inactive if its response is more than 10% above or below nominal, but nonuniformities in that range can be expected in the remaining elements. It is usually necessary, when interpreting CCD images, to correct these non-uniformities by means of a calibration image taken against a neutral background.

Digitisers

The digitiser itself acts on the signal generated by the optical sensor and, even if the digitisation characteristics influence the choice of the sensor, its design is relatively straightforward. Commercial A/D converters can be found in a variety of sampling speeds going from several MHertzs, suitable for video input rates, to a few tens of samples/sec, and with digitisation accuracies from 1 to 15 bits. They are usually reliable and their performance corresponds roughly to the one stated in their data sheets.

The source of error most commonly associated with A/D converters is nonlinearity in the decision levels for digitisation. It is usually assumed that the full input range of the converter is divided into equal voltage intervals, each of which is assigned a consecutive code number. If the limits of these intervals are not really equidistant, systematic errors result in the interpretation of the data. Linearity is usually reported in data sheets as a mean value (about 1/4 to 1 least significant bit) which may mask substantially larger non-uniformities in certain ranges of values. A common problem, almost never reported in the manufacturer specifications, is a concentration of nonlinearity near the transition of the high order bits.

ERROR ANALYSIS

The number of reliable bits that can be extracted from a digitisation system is a function of the signal/noise ratio in the sensor. Besides, the digitisation process itself is equivalent to an extra source of noise in the final signal. It can be shown that, if a signal with a total peak to peak range P is digitised in n levels, the equivalent rms noise introduced by the quantisation is

$$R = P/n*sqrt(12).$$

A way to design a digitisation system would be to choose the number of levels so that the quantisation noise is equivalent to the noise introduced by the source. Thus, if the sensor peak to peak signal to rms noise ratio is S/N db, we could equate

$$S/N = 20 \log(P/R)$$

which, expressed as a number of bits for the digitiser, gives

$$B = 0.17\ S/N-1.8\ . \qquad (1)$$

As an example, the digitiser for a typical vidicon camera with a S/N of 40 db would be designed with 5 bits.

The reliability of the last bit in a system designed in this way is rather low. Consider the 40 db example; the saturation signal is only 100 times larger than the rms noise of the source, while the interval is divided in 32 different intervals by the digitiser. Each interval is then equivalent to only three standard deviations and, assuming white gaussian noise, the signal will fall outside the "right" interval in 10% of the cases. In practice, a good rule is to take the decision band for the digitiser as 6 times the rms noise, which contains the true value 99.7% of the time. This criterion halves the number of levels and decreases by one the number of bits given by the formula above. The depths given in table 2 are computed in this way.

The digitisation accuracy can be improved by averaging several images; the amplitude of the averaged noise decreases proportionally to the square root of the number of images. It is a curious fact that, under some circumstances, the final noise can be reduced by the averaging to a level below the limit of quantisation noise. The condition for this to be possible is that the digitisation intervals be small with respect to the rms source noise. Under these conditions, different realisations of the signal are spread over many different quantisation intervals; the average is determined not only by the interval assigned to each realisation but by the frequency with which each interval appears. In fact, the digitiser is then acting as a histogrammer of the noise distribution and computing the average value from the histogram. In this way it is theoretically possible to increase indefinitely the accuracy of the measurement. Note, however, that this improvement can be fairly expensive; increasing the accuracy by an order of magnitude implies

averaging 100 consecutive images. Some vendors offer special hardware to do the averaging while the image is being acquired, but the price is normally high. If time is no problem, it is possible to do the averaging in the acquisition computer. However, the amount of core storage needed to accumulate the intermediate results is larger than that required for the final image. As an example, the values resulting from accumulating one hundred 8 bits/pixel images can range up to 25,500 and have to be coded in at least 15 bits/pixel. Finally, note that a real time system that has to be averaged a hundred times is not real time any more.

The option of being able to improve the S/N ratio by averaging forces the designer to overdimension the digitisation system. We have seen above that a digitiser designed according to equation (1) has decision bands which are 3 times larger than the rms noise amplitude. The requirement for averaging is that the intervals be small with respect to the noise and, therefore, a higher number of bits has to be provided. In practice increasing the number of bits by three above the limit given in equation (1) gives intervals which are about half the rms amplitude and which are enough for averaging.

In summary, equation (1) gives the design point for the number of bits that should be used in a digitiser so that the quantisation noise does not add appreciably to the input noise. Of these bits, the last one is unreliable 10% of the times, and the next one, 0.3%. In practice, systems should be designed with three bits on top of limit (1) to allow for improvement of the S/N ratio through averaging. In the 40 db case used above as an example, the minimum limit was 5 bits, a good design would include 8 bits, and the number of reliable bits in the absence of averaging would only be 4. Commercial systems are designed with ample margins above the S/N level of their sensors, in such a way that averaging is normally no problem. As a consequence, the number of reliable bits that they are able to offer bears little relation with the announced output of the digitiser.

Shading

Closely connected with the noise and accuracy characteristics of an image acquisition system is the uniformity of illumination in the camera and in the whole optical system. A theoretical limit for off-axis rays in a general optical system is that illumination decreases with the fourth power of the cosine of the angle that they form with the optical axis (6). In a vidicon with a 10x10 mm target and a 50 mm standard macro lens, the maximum angle is 8 deg. and the shading is only 5%, but the same lens in a CCD scanning camera with a 30x30 mm target (or in a 35 mm photographic camera) has a theoretical shading of 30%. In practice, shading is increased by vignetting, which is the blocking of some light by the entry pupil to the system, and is 1.5-2 times

higher than the theoretical limit. Shading in the illumination system used for the experiment is cumulative to that introduced by the camera.

Shading can have a significant effect in the average S/N ratio of an acquisition system. Assume that the illumination in the border of an image is only 50% that in the centre. The image acquired by the camera would have to be divided by the illumination distribution and the maximum signal possible in the borders would only be half of the theoretical saturation level for the camera; the signal to noise ratio would be reduced by a factor of two in those areas, and the number of effective bits would be reduced by one.

CALIBRATION

Calibration is an important part of any experiment, and image processing is no exception. The system to be calibrated includes the experiment itself, the illumination system, the camera and the digitiser. Of all these elements, the only one that can be considered more or less stable is the digitiser and, once it has been characterised initially, its decision levels can be taken as more or less constant. Cameras drift, most notably with temperature, but also with age. In extreme cases like SbS vidicon targets, this drift can happen in a matter of minutes. The importance of giving an adequate warm-up time to the electronics and to the instrumentation cannot be overemphasised. Even a good quality TV monitor cannot be considered stable until it has been connected for at least half an hour. Some devices have long time memory effects. The cathode of a photomultiplier (or a SIT target) that is exposed to a high light intensity can show an increased dark current for several hours or even days. Non-uniform sensitivity (shading) is a problem not only in CCD arrays. The sensitivity of a vidicon tube can change by almost 20% throughout the target. In this case the change is usually smooth, and may not be apparent to the eye, while in a CCD it is usually distributed randomly and appears like a peppery or streaky structure in the image. We have shown above that the shading by the optical system, even if it is smooth and not visually obvious, can be very important quantitatively. Remember that, from the calculations above, a 50% shading is present in most normal snapshots, slides and commercial TV programs.

When photography is used as part of the system, there is a whole new set of possible nonuniformities. Commercial photographic emulsions are fairly uniform but they change from batch to batch. Their response depends strongly on the development method. It is strongly recommended that critical pictures be developed by the experimenter as carefully as any other step in data reduction. Even if a uniformly exposed negative is developed carefully the result is not a uniform density. There are edge and adjacency effects in development that are almost impossible to avoid and that introduce variations in optical transmissivity of the order of 20%.

Photographic film is nonlinear and its gamma changes among batches and depends on the development method and on the particular batch of chemicals being used. Any sequence of pictures that is going to be used for quantitative measurements should include at least one picture of a neutral gray card to calibrate optical shading and a standard gray step tablet to calibrate film response. Good reviews of photographic image structure and development effects can be found in (2,3).

Normally it is possible to calibrate the whole optical system, including illumination and camera shading, with a single picture of a neutral background. In some cases, calibration can be more complicated. Consider a simple case of laser tomography in which a laser sheet excites fluorescence in a dye. The intensity of the fluorescence is proportional to the local intensity of the beam and to the local dye concentration. The problem is that the beam intensity changes in part due to the nonuniformities in the lighting arrangements and in part due to absorption by the dye. The obvious way of calibrating the optical system by taking a picture of a uniform dye concentration fails because absorption by the dye also contributes to the change in intensity. An easy remedy would be to work with a dye concentration low enough that the absorption be negligible, but that usually means that the fluorescent intensity is too low to be detected. The solution is to take calibration pictures with at least two different uniform dye concentrations and extrapolate to infinite dilution. Such methods are common in colorimetry.

DISPLAY

The final step in most image processing work is to display the image. Strictly speaking this is not necessary. It is possible to measure, for example, concentration statistics from a set of images without ever displaying them but, in practice, seeing at least the intermediate stages of the processing is essential in detecting errors and in guiding data reduction.

The ideal solution is to buy any of the good quality display systems available commercially. They consist of a high speed buffer that holds the image being displayed, a colour (or gray level) translation table and a monitor driven by the output of the table. Colour is desirable but not essential and can sometimes introduce artifacts by assigning completely different colours to neighbouring pixel values and creating artificial boundaries. If colour is available, the colour translation table is important. The user should be able to load it with randomly chosen colours from a wide palette. Some systems work by providing different buffers for each of the primary colours and mixing them only at the output stage. These system are designed to provide colour separation in graphic arts but, in scientific work, they waste storage and processing time and should be avoided if possible. Other

options include hardware zooming, panning, graphics, etc. All these are very convenient and should be weighted against the extra cost that they involve. Finally there are image processing systems that include with the monitor special purpose hardware to do a variety of image processing operations, usually much faster than any general purpose computer. Again, these systems should be considered with respect to their cost, which is generally high. It is sometimes advisable to try digital image processing with a simple system before getting into the expense of a full blown special purpose machine.

In fact, if output speed is not very important, it is possible to display images without special purpose hardware. A standard character printer can display about 10-15 gray levels by choosing denser characters for darker pixels, usually helped by overstriking. An all point addressable (APA) matrix printer can do the same thing by using half-toning techniques; a 4x5 "pixel" can be programmed to represent 21 different gray levels. Finally a pen plotter can be used if the image is seen as a relief map and represented with any of the available three dimensional graphic perspective packages.

All these methods are limited to relatively small images (less than 100 pixels wide) and are very slow (several minutes per image). Moreover, they lack the interactive capability of pointing and modifying an image in something close to real time. Speed and interactivity are very important in image processing. They can make the difference between successful processing and a long fruitless iteration. However, the casual user of DIP should consider the possibility of using the printing techniques mentioned above and, above all, should not be deterred from using image processing by the lack of more standard equipment.

A final point to be considered is hard copy. The easiest way to make a copy of the image in a monitor is to position a photographic camera in front of the monitor and take a picture. There is usually some distortion due to the curvature of the screen but it can be controlled by using lenses with a long focal length. There are special monitors with very flat screens and little distortion, designed specifically to be photographed, and which are normally found in the form of "cameras" containing the monitor, a photographic film holder and usually some exposure control circuit. These devices are becoming fairly inexpensive and the quality of the copy is good, but the cost of the film is normally too high for everyday work. Other copiers, ranging from electrostatic plotters to ink jet printers work by tapping directly the video signal from the monitor. Again the best guide is the manufacturers specifications.

CONCLUSIONS

The discussion in these notes might seem at moments as a warning against the use of digital image processing. This is not so. In fact, a similar work on the use of hot wire anemometry in fluid mechanics or of the telescope in astronomy might be full of similar caveats, while it is obvious that both are indispensable instruments in their fields.

There is a tendency to expect that any new experimental technique would solve all the problems in a given field with essentially no cost but the financial one. This is almost never true. Digital image processing is a new instrument and, as such, it can be used to solve many problems that could not be solved otherwise. However, it is only an instrument and it has to be designed with the same care and used with the same caution as any other instrument.

REFERENCES

(1) "IEEE Standard Digital Interface for Programmable Instrumentation". ANSI/IEEE Std 488-1978, Inst. Elect. Electr. Eng., NY. (1980)

(2) "SPSE Handbook of Photographic Science and Engineering", W.Thomas,Jr. (ed.). Wiley-Interscience, N.Y. (1973)

(3) "Kodak Plates and Films for Scientific Photography", Kodak P-315, Eastman Kodak Co., Rochester, N.Y. (1973)

(4) A good survey on electro-optical sensing devices can be found in: "Manual of Remote Sensing", Vol.1, Second Edition, Ch. 7-8. R.N.Colwell (ed.). American Society of Photogrammetry, Falls Church, Va. USA (1983)

(5) L.R.Rabiner & B.Gold, "Theory and application of digital signal processing" Ch.5, Prentice Hall, Englewood Cliffs, N.J., U.S.A. (1975)

(6) M.Born & E.Wolf, "Principles of Optics" Ch.4, Pergamon Press, London, (1959)