The challenge of picture processing

G. Groh

Introduction

About one-and-a-half thousand million years ago a biological evolution began that led to the origin and the development of the visual system. Some microorganisms became capable of distinguishing between light and darkness and learned to react accordingly. This accomplishment greatly improved their chances of survival. An extremely primitive kind of visual scene analysis had proved its usefulness. The rest of the development is well known. Arrays of sensors with increased processing power appeared and finally — at a rather late stage — the eye with its image-forming lens was 'invented'.

Technological evolution followed a different course. Image formation by means of mirrors and lenses and picture recording by photography were the starting points. Later some methods of processing were introduced to compensate to some extent for the often poor quality of the products of the early systems. Picture restoration by manually retouching photographs is the best-known example. Other methods aimed at an improved presentation of pictorial information by manipulating the image. Edge enhancement, which gives an apparently sharper or clearer picture, is one example. The latest stage has been the advent of more 'intelligent' procedures, such as coding for data compression, feature extraction and classification, pattern recognition and finally automated picture evaluation, all strongly stimulated by the increasing capabilities of digital computers.

Thus, the technical discipline of picture processing has three different origins. The first is optics and photography, with its fairly long tradition. The second is television, whose refined analog electronic facilities allow the implementation of many methods of picture manipulation in real time. The third and most recent one is digital data processing, which enables artificial intelligence to be used in picture processing.

These different schools of technological development emanating from different scientific disciplines have not yet been successfully merged together. Compared with the biological competitor, technical picture processing is still in its infancy. That is the challenge!

I shall return to the biological world in the last section of this introduction. For the time being I shall sketch briefly some of the major applications and methods of modern picture processing. Obviously, this cannot be a complete description of the present state of the art [1]. The emphasis will be placed on those areas that are related in some way to the topics dealt with in the other articles in this issue.

First, I shall look at some of the methods of picture processing from the viewpoint of applications. This will demonstrate the fairly long tradition of picture processing and will also indicate the motivation for the considerable research effort being put into this field. Secondly, I shall describe some of the basic principles underlying the methods used. Thirdly, I shall discuss some of the techniques used to perform these operations. Each of the different views will confront us with open questions and reveal some seemingly intractable problems. These are challenges for future research.

Applications of picture processing

There are many areas of human activity where picture processing is useful or necessary. Instead of trying to present a comprehensive but perhaps somewhat trivial list, I shall focus on two major areas of application that are both important in themselves and representative of many others: medical diagnostics and industry.

In present-day medical diagnostics wide use is made of X-ray pictures, or radiographs. These are 'shadowgrams', produced by exposing the patient to irradiation from a point-like X-ray source and recording the projected absorption pattern ('radiation relief') on a photographic film. The absorption differences vary over more than three orders of magnitude. The radiograph therefore represents a grey-tone picture with an enormous dynamic range in optical density, which can only be reasonably covered by means of special double-
coated films. Closed-circuit television, coupled to X-ray image intensifiers, which is increasingly used for certain types of examination, only shows up to about 50 grey levels, however. It is primarily the display equipment that restricts the use of advanced electronic systems to applications with less exacting requirements.

This physical limitation, however, can be easily overcome by means of a simple type of processing called 'dynamic windowing'. Only the part of the intensity range of interest to the human observer is selected and displayed with the full grey scale of the equipment. Two control knobs allow him to scan the full range and to define the 'width' of the 'window'. One application in which this technique is now used is in displays for computerized tomography [3].

Image formation in radiography is often limited by photon noise. The appearance of the noisy picture can be improved by applying spatial smoothing techniques. This corresponds to a slight defocusing of the image, which gives an averaging over the statistical fluctuations in the neighbourhood of each individual pixel. The spatial resolution of radiographs is limited by the size of the focal spot of the X-ray tube. The apparent sharpness of the picture can be improved by edge enhancement. This method was introduced as long ago as 1930. It was then called 'harmonization' and made use of photographic processes to superimpose the original picture and its slightly defocused negative [3].

When we consider the great density range of radiographs, it is obvious that coarse, irregular structures of low contrast may remain indiscernible. A solution to this problem is contrast enhancement by altering the distribution of grey levels in the picture.

Examples of applications of these techniques are given in the articles by A. Hoyer and M. Schindwein [4], A. Hoyer and W. Spiesberger [5] and L. H. Guildford [6] in this issue. Whereas the first two papers deal with general digital data-processing systems, the third discusses a new approach for real-time equipment, which is important in such applications as infrared imaging and television at low light levels.

Before discussing further the increasing refinement and complexity now found in the applications of picture processing, I should like to draw attention to a relatively new development, which is growing in importance but is sometimes overlooked in this context: image formation by picture processing.

In medical diagnostics a prominent example is computerized tomography, in which a rotating linear array of X-ray detectors is used to measure the absorption values in about 100 000 locations and directions of a cross-sectional 'slice' of a patient's body. The 'one-dimensional shadowgrams' thus obtained are processed so as to generate a two-dimensional picture of the local absorption coefficient in that slice. This is not the place to go into the rather complicated mathematics involved [2]. It will be useful, however, to look briefly at the ideas underlying this approach.

Radiology deals with the imaging of three-dimensional objects by means of X-radiation. Physics does not offer any advanced image-forming devices for X-radiation. All we have are highly primitive shadowgrams, showing a superposition of all the layers at different depths in the patient. Without a priori knowledge, location in depth is not possible. Owing to the superposition of so many structures, even the considerable density range of an X-ray film is not sufficient for recording and resolving the tiny absorption differences in soft tissues, something that is of primary importance in early cancer diagnosis.

The solution offered by computerized tomography is a two-step process of image formation. First, a well-defined set of data is generated, in accordance with the physical measuring possibilities and boundary conditions. Then the desired final image is formed by post-processing. As described earlier, potential limitations of the display unit are overcome by picture manipulation.

Other medical examples of image formation by picture processing are tomosynthesis [7] and coded-aperture imaging. A very recent version of the latter is described in the article by E. Klotz et al. on flashing tomosynthesis [8]. In this case imaging of the three-dimensional object is performed by simultaneously 'flashing' an array of X-ray tubes, irregularly distributed in a plane above the patient. The resulting superposition of images, called the coded picture, is not directly comprehensible but contains information on the depth. Only after appropriate processing can pictures of slices at different depths — the tomograms — be constructed.

Two-step image formation using picture processing is also gaining acceptance in other medical applications, such as ultrasonic imaging [9] and spin imaging [10]. The latter utilizes nuclear magnetic resonance for characterizing tissues. Similar principles have already established a certain tradition in nuclear medicine, for example in generating patterns of the distribution of isotopes in the human body. A much older and well-known example in another field is 'synthetic-aperture radar' [11], which again uses picture processing to form a radar image with vastly increased spatial resolution.

The next step on the ladder of increasing refinement is automated recognition of patterns in a picture. The most prominent sample for many people will be machine reading of printed or even handwritten docu-
ments by means of optical character recognition [16]. Other applications are found in medical diagnostics [19]. More stringent demands, however, are nowadays found in industrial applications. Examples are the inspection of printed-circuit boards, mask inspection in the production of integrated circuits, and also non-destructive testing of materials and semifinished products by means of X-radiation [14], with a view to detecting flaws such as pores in welding seams, cracks, porosities and cavities in castings, or irregularities in the fabric structure of rubber tyres. In general, quality control by automated inspection of pictures — which are generated by infrared radiation, visible radiation, X-rays or ultrasound — is gaining increasing importance now that many industries manufacturing 'critical' parts for motor cars, aircraft, nuclear power plants, etc. are insisting on an extremely high reliability for their products.

The hope is that automated inspection by means of computerized pattern recognition will prove to be more reliable and more objective than a wearisome visual inspection by people. As yet, however, no general solution has been found.

The severest requirements of all come from visual scene analysis. This is an important function in robotics, which is mainly used in assembling products, sorting goods, controlling vehicles, etc. In these cases an 'intelligent' interpretation of a three-dimensional scene under different kinds of illumination will ultimately be required. This means that the machine has to recognize the shape, location and orientation of a two-dimensional object. It has to distinguish real parts of an object from shadows. It must not be misled by printed patterns on the surface of the object. The spatial relationship of parts that overlap in one projection have to be recognized and interpreted. And all these processing procedures have to work in real time on sets of grey-tone pictures showing different perspectives.

This applies to general-purpose machines for automated assembly work. Fortunately, there are other applications where the objects are flat (essentially two-dimensional) and where uniform illumination from behind can be used to generate black-and-white pictures. These can be unambiguously described by features of their contours, which considerably reduces the amount of data needed and makes real-time processing possible. The possibilities and potential of visual scene analysis in this area of applications are described in the article by P. Saraga and J. A. Weaver [15], who have successfully made an experimental visually controlled pick-and-place machine.

The examples given in the foregoing represent only a very small fraction of the full range of existing and potential applications of picture processing. The poten-
tialities are so promising that they are motivating a vast research effort all over the world. In Japan, for example, a huge research programme on 'Pattern Information Processing Systems' (PIPS) is being sponsored by MITI's Agency of Industrial Science and Technology [14]. This project was started in July 1971 and estimates put its total costs at about 25 thousand million yen (about 100 million dollars) by 1980. Major participants are the leading Japanese companies like Mitsubishi, Toshiba, Hitachi and Fujitsu. In the United States of America a government-sponsored research program on 'Image Understanding' is in progress, which has an approved annual budget of three million dollars in the years 1972 to 1981 (this does not include military applications).

Some principles

Various different methods and techniques are used today for picture processing, depending on the different disciplines in which the picture originated. This
does not matter and can even have its advantages, as long as the application-oriented approach is adopted and the best combination selected in each case. Picture processing requires an interdisciplinary attitude. Neither optics nor television electronics nor digital data processing alone can offer general solutions, as is clearly demonstrated by the articles in this issue. Nevertheless, there does exist a set of fundamental operations, which are used again and again. I shall try to illustrate this with a few examples that all relate to two-dimensional grey-tone images.

Mathematically, two-dimensional grey-tone images can be described by a brightness or intensity function \( I \), whose value varies with the location \((x,y)\) in the image plane:

\[
I = f(x,y). \tag{1}
\]

The black-and-white ('binary') image is a special case in which \( I \) takes one of two possible values at each point \((x,y)\).

Some simple primitive operations are negation (i.e. black changes to white and vice versa) of the images and addition of several images. Others are translation, rotation, scaling and distortion or stretching. Their usefulness and potential is demonstrated in the article by Klötz et al. [1] and — indirectly — in the article by Saraga and Weaver [18,19]. Whereas in scene analysis these operations are mainly used for searching and comparison and have to be supplemented by more refined operations, in tomosynthesis only translation, scaling and addition are necessary for performing the decoding process.

Integration over or addition of a sequence of noisy pictures of the same object taken at different times has the effect of reducing noise. Edge enhancement can be obtained by spatial differentiation. This can be approximated by adding the negative of the blurred image to the original one, as we have seen earlier.

Here a new operation comes in: blurring or defocusing, which can be described as weighted spatial integration over the neighbourhood of each pixel. This is easy to understand by considering the simplest of all images, i.e. a very small white spot on a black background. Defocusing will make this spot larger and a smoother transition will take place from black to white. The new intensity profile is called the point-spread function of the system that causes the blurring. A more interesting picture obviously consists of many pixels; its blurred version is the superposition of all their point-spread functions.

Of course, a pixel is a mathematical fiction that does not exist in reality. Every real image-forming system has a limited resolution, which is determined by the width of the point-spread function. For this reason it is permissible, without losing too much information, to scan a picture with a small spot line by line, as in television, to convert it into a time-sequential signal, or to describe the picture as a two-dimensional matrix of discrete individual pixels, to enable it to be processed by a digital computer. However, the scanning or pixel sampling has to be fine enough to prevent the raster of scanning lines or pixels from showing up.

There can be no doubt that scanning and (two-dimensional) spatial sampling are the most important basic operations for converting an image into a time-sequential signal or a list of computer data and vice versa. More details on conversion by scanning will be found in the article by Guildford [6].

The digital computer needs an additional operation. The original analog intensity values have to be quantized, i.e. converted into a finite number of grey levels. Again the quantizing operation has to be fine enough to avoid 'artefacts'.

Further processing of signals or sets of data generated in accordance with these basic conversion operations is easily performed. In particular, nonlinear distortions of the grey scale, as in thresholding methods (for generating binary images) and dynamic windowing methods (to stretch and clip the grey scale) are simply a question of using conventional electronic circuits or of rearranging data.

In most pictures the spatial context of the pixels is of decisive importance. This is taken into consideration by the local operator, a small window or 'keyhole' of say \(5 \times 5\) pixels, which can be deliberately moved over the whole picture. The operations described earlier, or others like spatial gradient operations, are carried out only within the area of this window. Examples of this technique are given in the article by Hoyer and Spiesberger [5].

An important characteristic of a picture is the distribution of the probability of occurrence of the grey levels, the grey-level histogram. The manipulation of the grey-level distribution yields impressive results. Examples will be found in several articles in this issue [4–6].

Finally, spatial filtering should be mentioned. The basic operation is the two-dimensional Fourier integral of the image function \( I \):

\[
\mathcal{F}(I) = \iint f(x,y) e^{-2\pi i (\nu x + \mu y)} \, dx \, dy.
\]

The new variables \( \nu \) and \( \mu \) are called spatial frequencies and correspond exactly to the well-known signal frequency in the time domain.

Spatial filtering means manipulation of the spatial-frequency content of the image. For example, suppression of the high frequencies (lowpass filtering) reduces the sharpness of the picture and causes de-
focusing. Suppression of the low frequencies (highpass filtering) results in a sharper picture.

Although far from complete, this list of basic image operations will serve to illustrate the potentialities. They can be carried out in many different ways. The same or similar procedures as have been outlined above are used for feature extraction. The general principle is firstly to enhance special features (e.g. edges, contours, local maxima, spatial frequencies) in the picture itself or in the spatial-frequency domain, and secondly to apply a threshold or masking operation to suppress the rest of the image.

In most cases the extracted features themselves form a two-dimensional pattern, i.e. an image which contains less information, but usually still too much.

In the next step of data reduction some of these features are numerically described, resulting in data sets that are called 'features'. The nature of the feature of interest depends to a great extent on course of the particular application. In medicine the doctor may only be interested in the shape of the heart contour or in the area it encloses. If all the other information about location or orientation is irrelevant, one ends up with a small set of data or even with a figure. In mammography all that is required for further investigation is the location of bright spots in the mammogram that betray the possible presence of micro-calcifications. In non-destructive testing of the rubber tyres of motor cars, deviations from the periodic structure in the embedded fabric have to be detected and localized. In visual scene analysis an enclosed contour may have to be characterized further to see whether it is perhaps identical with the stored model of the object. If it is, only its orientation and the location of its centre of gravity may be of further interest.

These few examples should provide sufficient illustration. The number of applications requiring modified approaches is overwhelming. This demonstrates the main problem and the difficulties of more advanced picture processing, namely the characterization and extraction of useful features from images. Some general principles are known, but there is as yet no complete theory from which solutions for all conceivable problems could be derived, at least in principle. The classification process is better understood; it does at least have some underlying theory.

In nature, biological evolution is not based on deduction. Species of plants and animals are to be considered as self-organizing systems. Mutations that lead to a better chance of survival will remain, others will fade out. This holds in particular for the 'intelligent', i.e. image-understanding part of the visual system. During its evolution numerous modifications of the 'hardware' have been tried and selected for their 'usefulness' in the struggle for survival. Even during the life of the individual creature, the system and especially its 'software' is further trained and improved.

In the world of technology there is nothing comparable. The wishes and dreams are as old as computer science. As early as 1953 C. E. Shannon [17] formulated a set of challenges, from which I can quote in this context:

'Can we organize machines into a hierarchy of levels, as the brain appears to be organized, with the learning of the machine gradually progressing up through the hierarchy? . . . Can manipulative and sensory devices functionally comparable to the hand and eye be developed and coordinated with computers? . . . Can more satisfactory theories of learning be formulated? . . . How can a computer memory be organized to learn and remember by association, in a manner similar to the human brain? . . .'

These questions could have been asked today, a quarter of a century later.

Nevertheless, considerable progress has been made in artificial intelligence. The article by Saraga and Weaver in this issue gives one example [18]. The articles by T. J. B. Swanenburg [18] and E. H. J. Persoon [19] describe some recent experiments on self-organizing memories and recognition systems. Although the class of objects in the pictures is rather restricted, the feasibility of the procedures is clearly demonstrated and there is every expectation that they can be extended to more general pictures.

**Practical realization**

By far the most important progress has been made in the technology, i.e. in the hardware. In 1949 W. S. McCulloch [20] made the picturesque remark that a computer with as many tubes as man has neurons (10^10) would require the Empire State Building to house it, the Niagara Falls to power it and the Niagara River to cool it (the human brain dissipates some 25 watts!). Today we are starting with Very Large Scale Integration (VLSI) which promises by 1985 to achieve integration of 10^8 to 10^9 gates on a thin silicon chip of one square centimetre! If closely packed, 10^9 neuron-like gates would require the volume of a shoebox. The design and construction of an intelligent visual system like the one that forms part of the brain has become technically feasible, at least in terms of volume.

---

The real bottlenecks, however, still exist. We do not know how to organize such an enormous number of logical functions, that is to say the architecture of such a system and most of the more advanced algorithms are as yet unknown. We even have difficulties in feeding the data flow from a high-resolution television camera into a computer in real time [21].

One of the reasons is that both our television and our computer systems are organized time-sequentially. This is poorly matched to pictorial information, which is essentially parallel. For example, the standard television picture of about 300,000 pixels (= 500 x 600) each consisting of 64 permissible grey levels (six bits) contains as many as 1.8 million bits of information. The elementary operation of just adding two pictures by sequentially passing these pixels through an adder circuit, requiring only a hundred nanoseconds per operation, will take in all 30 milliseconds, which corresponds roughly to the normal frame rate. More refined operations will require much more time. The only solution to this problem is parallel transmission and processing of pictorial data.

There is one approach which is by its very nature parallel, i.e. optics. Many linear operations, like spatial filtering, can be easily carried out in real time on large amounts of data. The only limitation in this respect is the velocity of light. Klitz et al. demonstrate in their article in this issue [8] the simplicity and elegance of optical decoding. Another very successful approach is used in synthetic-aperture radar [11]. Numerous system proposals for advanced applications, using coherent light and holography for pattern recognition, have also been discussed in the literature [22]. They all have the advantages of parallelism, but fall short in three important aspects:
- the class of image operations that can be carried out is too restricted; they are mainly linear operations;
- these operations cannot be changed quickly under program control, i.e. they resemble hard-wired filter functions;
- optical processors are not yet compatible with electronic or digital systems, which could give the desired flexibility and make other operations possible, for example Boolean operations.

This situation would soon change with the appearance of new interface components for input, output and filter modification. Then, of course, the analog or digital electronic circuits will still have to be structured in parallel. Otherwise, most of the time will be lost in parallel-to-sequential conversion of the signals and vice versa. The absence of such interfaces and the incompatibility of the subsystems are the main reasons for the present bottleneck. We can make picture-processing systems in one of the three basic technologies, but can seldom come up with combined systems that utilize the specific advantages from each system.

By definition the general-purpose digital computer is very flexible and useful for a wide range of operations. It is therefore a most excellent tool for investigating new algorithms, methods and applications, and for testing the procedures by simulation. For practical use it is usually much too slow, too difficult to handle, too clumsy and too expensive.

The trend is towards dedicated digital processors, which offer just the right degree of programmability and range of operations required by the special applications. Growing use is being made of the parallel architectures that have become economically feasible in modern microprocessor technology [23]. For input and output these dedicated processors are interfaced with analog electronic components, such as camera tubes and cathode-ray-tube displays.

The challenge

The continual challenge, however, is the biological visual system. It shows us what is in principle feasible. Perhaps the study of its functioning may stimulate our creativity and lead to novel, unconventional approaches. I shall therefore briefly summarize some of its more remarkable features.

In the human eye, picture processing starts at the retina, which consists of $1.25 \times 10^8$ receptors (94% rods for low light levels, black-and-white, and 6% cones for colour and daylight). They feed only $10^6$ nerve fibres, which transmit the information to the brain. This considerable data reduction takes place mainly in the periphery, where up to several thousand receptors (rods) are somehow connected so as to jointly transmit their preprocessed information along only one nerve fibre. The principles of this data compression still remain a closed book. The greatest attention is paid to the two thousand cones in the rod-free fovea centralis. Each cone has a nerve fibre of its own. Their information is projected on to a disproportionately large area of the visual brain without substantially changing the spatial arrangement of the pictorial information.

In the main area of visual reception around the fovea centralis (corresponding to a viewing angle of about two degrees) some $6 \times 10^4$ pixels are resolved,

---


each capable of identifying about 250 grey levels (i.e. eight bits). The maximum information rate is 10 Hz at an illuminance of one microlux, but goes up to 60 Hz at one lux. This corresponds to a total flow to the brain of $6 \times 10^4$ (pixels) $\times 8$ (bits) $\times 10$ (Hz) or $5 \times 10^6$ bits per second. These five million bits per second are then further processed and reduced. The rate of perception of the visual system is less than 300 bits/s! As we know, for example from experiments on reading, the maximum total processing rate measured in terms of the conscious output, i.e. what finally gets through to our consciousness, is only about 50 bits/s. This means that the main part of the visual information (not even taking the periphery of the retina into account) is reduced by a factor of more than $10^5$ without any loss of the relevant ‘useful’ information. That is the miracle and the challenge!

Some methods that we use in technical systems are used by the visual system as well. The best known is Mach’s effect, which corresponds to an edge enhancement. But in addition the eye has a remarkable capacity for adaptation, even to systematic distortions or imperfections of the imaging system. For example, after a training phase of a few weeks it can even compensate for a 180° rotation of the images it receives (i.e. an upside-down world) or correct for severe chromatic aberrations.

Another interesting feature, whose effect on picture processing is not understood, is the irreducible tremor of the eye. It consists of irregular small movements of 30 seconds of arc, on which are superimposed slow drifts and saccadic movements of up to five minutes of arc. All we know is that without these irregular movements only moving objects would be observed, while stationary parts would gradually become invisible.

Is this just to compensate for a shortcoming of the light receptors, which apparently cannot detect light of continuously equal intensity, or is an additional preprocessing phase involved, which somehow takes the environment of every pixel into account statistically?

Technologically we process pictures with the aid of a computer, which is interfaced to a television-like receiving system. Most procedures are time-sequential. We are only at the very beginning of the introduction of parallel processing methods. Biologically, we have the fully parallel organized brain, part of which is coupled to the retina of the eye directly and in parallel. There is no real interface, since the ‘format’ of the signals does not change. ‘Computations’ for picture processing begin right at the detector, take place in some way or another even during transmission, and are finally completed in the brain. This multilevel processing is facilitated by a uniform structure of almost identical components — neurons. Embryologically and functionally the retina is part of the brain.

It would certainly be too naive to try to copy the biological system, whose functioning we do not even understand. On the other hand the differences between the most successful biological system and our rather mediocre technological approaches are very fundamental and most striking.