3

Computer-generated cartoons

DON PEARSON, E. HANNA AND K. MARTINEZ

Introduction

The cartoon or line drawing is a very economical method of portraying a person or scene. Studies of the criteria used by humans in the recognition of faces have indicated that simple geometrical considerations (the distance between the eyes, the width of the mouth, and so on) predominate[1]; these can be adequately conveyed in a cartoon. Hand shapes can also be satisfactorily represented in this way; for example, the Royal National Institute for the Deaf publishes the standard manual alphabet used for finger-spelling in cartoon form[2].

We have become interested in generating cartoons electronically, as a way of transmitting moving images over the telephone network[3–7]. Since the telephone network has been designed to carry low-frequency speech signals, it can take up to several minutes to transmit a single frame of a conventional television picture. Experiments have shown, however, that enough small cartoon images can be transmitted every second to create the illusion of movement. As these images have only two levels of luminance, it is possible to use a very different code similar to that used in transmitting documents over telephone lines by facsimile. Such codes, termed 'run-length' or 'relative-address' codes, send the image information as a succession of numbers; each number represents the distance of a black/white transition from a neighbouring transition[8].

One of the aims of our work, which we share with others [9–11], is to enable the deaf to communicate over distances by signing or lip-reading. Since sign language relies on hand shape, orientation, position and movement, together with facial expression, it can be adequately conveyed in moving cartoons. Another of our aims is to use the cartoon as a foundation or primitive on which to build up shading and colour[12]. This may allow moving colour images for video-conferencing to be transmitted very economically over the new digital telephone and data networks which are now being introduced.

There is a double constraint which these engineering aims place on the cartoon: it should satisfactorily portray the face and hands of the person concerned, but it should do so with as few lines as possible. The more lines there are, the more the information which has to be transmitted per image or per second in the run-length code; a consequently greater 'channel capacity' is required to carry the signal. The channel capacity needed for a broadcast-standard studio television picture is about 20,000 times that available on a telephone line; extreme economy is therefore needed in the cartoon.

Electronic generation of cartoons

The subject of how to draw an economical cartoon was considered in the 1950s and 1960s in terms of information theory[13], the idea being that cartoon lines should be placed where the object is difficult to predict and the line therefore carries the most information. However, the validity of this approach has been questioned[14]. It also presupposes a more basic operation needed to convert an image into a line drawing (of whatever kind). The prescription for this basic operation is essential in any electronic implementation.

Our first approach to electronic generation was to try to detect edges in the image. An edge is a sudden change in the luminance of the image (Fig. 3.1a). It seemed reasonable to suppose that the locations of these changes would be useful to the eye. To detect the edges, a computer can be programmed to search for a particular pattern of lighter and darker image elements, indicated as + and − signs in Fig. 3.1b. This is accomplished by scanning the

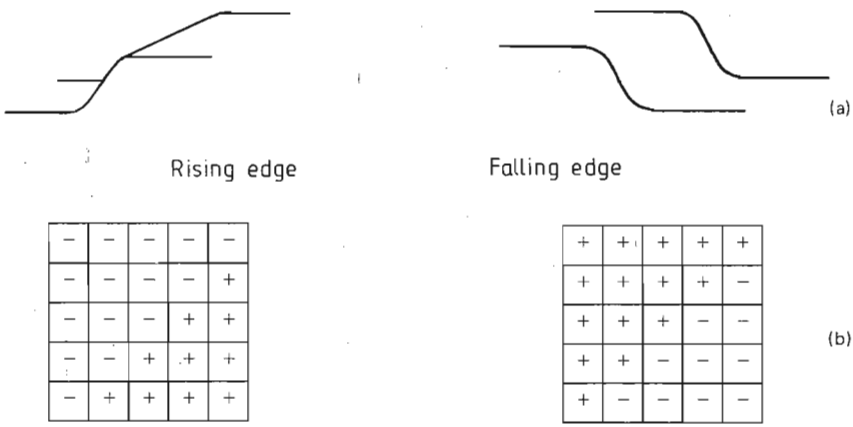


Fig. 3.1. (a) Luminance changes corresponding to an edge in the image.

(b) The pattern of lighter (+) and darker (−) image elements which a computer searches for in order to detect an edge at 45° to the vertical.

electronically sampled image using a 'window', shown in Fig. 3.1b as a 5 × 5 array of image elements. Since the edge can lie in any angular direction, this may have to be repeated with different patterns sensitive to different edge orientations.

After fairly lengthy experimental investigations which used many different published edge-detection procedures, we concluded that edge detection, when applied to the human form, does not succeed very well[4–6]. It tends to produce too many lines and at the same time to omit some key ones. Edge-detected images of humans look vaguely human in outline but have poor facial features. Hands held in front of faces are inadequately rendered. In consequence we attempted a new approach.

Theory of computer-generated cartoons

The new approach was to formulate the rule for drawing cartoons in the three-dimensional space of the object rather than the two-dimensional space of the image. It seemed to us, as a matter of intuition, that the important surfaces of the face and hands were the ones which fell away sharply from the line of sight of the television camera (or the eye of an observer).

If we imagine straight lines drawn from the lens of a camera in all conceivable forward directions, some will pierce the surfaces of objects in the field of view, others will miss them altogether, but a small subset (Fig. 3.2) will just graze the surface. The basic postulate of the theory[6] states that a cartoon point (usually forming part of a cartoon line) should be drawn in the image plane wherever a line grazes the surface of the object in this way[15]. On any given scan line in the image (Fig. 3.2) there are only a small number of these points (indicated as A, B and C) which have to be identified.

It is necessary to pay attention to scale[16] in this definition; the correct degree of conceptual smoothing has to be applied to the physical microstructure of the human form, so that the cartoon has neither too much nor too little detail. It is also helpful to provide a relaxation of the rather strict criterion of lines being tangent to the surface, to allow slight penetration of the surface by the straight line; this accommodates features like the edges of noses. In engineering it is wise to build in tolerances.

Implications and implementations

Since in practice the cartoon extractor is required to work with images and not objects, it is necessary to analyse the varieties of illumination falling on smooth surfaces defined by the postulate. If light is reflected to the camera at such surfaces in a consistent way, producing a well-defined feature in the image, it may be possible to locate the surfaces by an operation on the image.

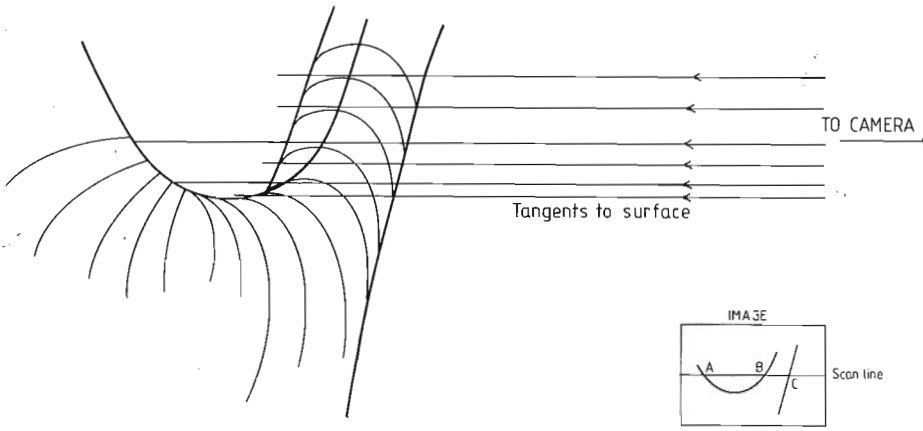
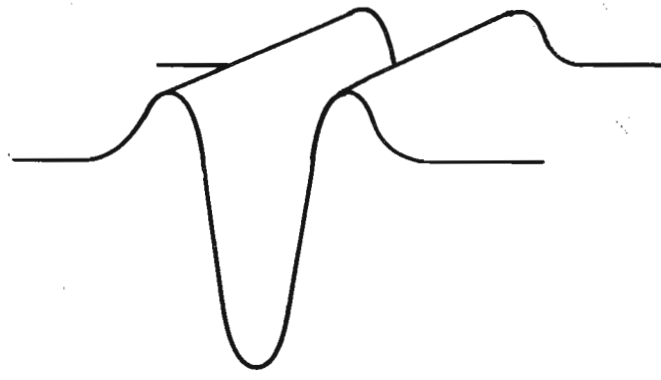


Fig. 3.2. The basic postulate in the proposed theory of cartoons identifies surfaces which are grazed by straight lines drawn from the camera. When the image of the object is scanned, the computer has to identify points A, B and C as cartoon points.

This analysis has been carried out, taking into account mutual illumination, that is light reflected from one surface to another before being reflected to the camera[6, 17]. It has been found that the image feature produced at a surface satisfying the postulate is a luminance *valley* when all local surfaces have the same reflectance (for example at the side of the nose or at the side of a finger held in front of the face); it is, however, a luminance step edge at an occluding contour seen against a background of different reflectance (for example, a hand held in front of a wall of greater or lesser lightness than skin). We have therefore argued that what is required is a feature detector with a primary response to luminance valleys, but possessing a secondary response to luminance edges.

Figure 3.3 gives a simplified portrayal of the type of detector required. Its action may be contrasted with that of the edge detector in Fig. 3.1. Like the edge detector, it searches for a particular pattern; in this case, it is a pattern consisting of a line of relatively dark elements in the image (represented by minus signs), surrounded by relatively light ones (the plus signs). This is what would commonly be called a valley in geographical terms. In practice the differences between the minus elements and the plus elements can be quite small.

Like edge detectors, valley detectors are defined on a window of image elements (5×5 elements in Fig. 3.3). The size of the window affects the ability of the detector to discriminate in favour of valleys and against other features (such as edges). The situation is rather like that of a radio receiver. Some radio



(a)

Fig. 3.3. (a) Luminance feature to which a valley detector is primarily sensitive. This may also be viewed as the required impulse response of a spatial filter, preceding a peak detector, which will be maximally sensitive to such a feature.

0	+	−	+	0
0	+	−	+	0
0	+	−	+	0
0	+	−	+	0
0	+	−	+	0

(b)

(b) Pattern searched for by the detector for a valley running north to south. The characteristic of the pattern is a column of slightly darker (−) elements seen against a surround of slightly lighter (+) elements. The elements marked 0 allow for some flexibility in interpretation.

receivers have sharper tuning than others, so that they can more easily reject unwanted radio stations with similar broadcasting frequencies to the desired station. The rejection of unwanted image features is carried out by detecting only those valleys whose depth is greater than a preset amount[18].

The 'tuning' of a valley detector gets sharper as the window size increases. Unless it is of infinite size, it will have some response to edges. So to produce a valley detector with a secondary response to edges, it is only necessary to make its window of finite size. This is fortunate since, as remarked earlier, the scale of the valley – that is, its width and length – must be matched to the size of the face or hand in the image for the best results.

In our actual implementation we have used certain refinements on the simplified valley detector given in Fig. 3.3. These are mainly to increase the speed, to discriminate against noise and to ensure that only the valley floor is identified. Figure 3.4 describes an actual 5×5 detector we have used[6, 17]. This was the one which generated the cartoons illustrated in the next section.

We are able to generate moving cartoons in real time from 64×64 element images at 6 frames/s using a 3×3 window and a 68000 microprocessor. We have also developed a simple parallel architecture which can process 256×256 images at 25 frames/s[19].

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o
p	q	r	s	t
u	v	w	x	y

a — y are picture elements

T_1, T_2 are thresholds

To test for a vertical valley through m:

if $((l - m) > T_1$ or $(n - m) > T_1)$

then

if $(f + k + p + j + o + t - 2(h + m + r)) > T_2$

and $(g + l + q + i + n + s - 2(h + m + r))$

$> (f + k + p + h + m + r - 2(g + l + q))$

and $(g + l + q + i + n + s - 2(h + m + r))$

$> (h + m + r + j + o + t - 2(i + n + s))$

then there is a valley through m.

Tests for a horizontal valley and for diagonal valleys have the same form, but the thresholds for the diagonal valleys are higher.

Fig. 3.4. Actual implementation of a 5×5 cartoon operator, comprising a valley detector as in Fig. 3.3, with elaborations.

Comparison of computer-drawn and human-drawn cartoons

Two examples of computer-drawn cartoons are shown in Figs. 3.6b and 3.7b, which were derived from the photographs shown in Fig. 3.5. These photographs were first converted into discrete electronic representations of 260×245 picture elements, each element having 64 grey levels. The electronic images were processed to produce the cartoons by use of the cartoon operator in Fig. 3.4. In addition, some uniform black shading was added for effect. This shading was produced by 'thresholding', that is, by identifying all the areas in the original grey-level images which were darker than a certain predetermined value. This is easily done electronically and does not add significantly to the amount of information which has to be transmitted.

To see how the computer's efforts compared with those of a human artist, we asked a professional cartoonist (Mr S. Wood, known widely as 'Woody') to emulate the computer. He too was given the photographs of the male and female subjects in Fig. 3.5. He was not shown the computer-drawn cartoons, nor any other examples of the computer's work, but the required style was described (as few lines as possible, with uniform black fill-in allowed). No indication was given as to where he should place his lines nor of the cartoon operation which the computer uses. The human cartoonist's renditions are shown in Figs. 3.6a and 3.7a. He also decided to caricature the two subjects, as shown in Fig. 3.8.

As non-artists, we were struck by the considerable similarity between the human-drawn and machine-drawn versions, while noting some differences.

Fig. 3.5. Original photographs of (a) female and (b) male subjects.

Fig. 3.6. Comparison of (a) artist-drawn and (b) computer-generated cartoons of the female subject.

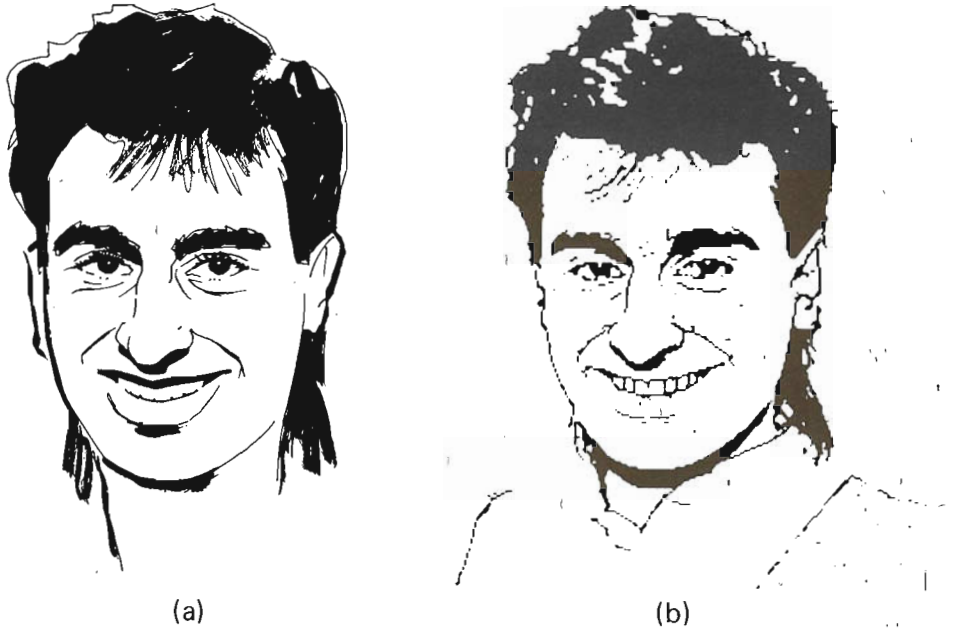


Fig. 3.7. Comparison of (a) artist-drawn and (b) computer-generated cartoons of the male subject.

Fig. 3.8. Caricatures by the human artist.

The main feature lines around the eyes, mouth, nose and chin are remarkably alike, although the human artist does not show the same detail in the teeth. The black shading in the hair, in the corners of the mouth and in the eyebrows of the man is very similar. The number of lines used by the cartoonist and by the computer is about the same for the face, excluding the clothing; however, because of the relatively low sampling density used in the processing, the lines in the computer versions have a jagged appearance. There are differences, too, in higher-level interpretation of the photographs (of which the human but not the computer is capable) as to what is and what is not of importance. For example, the woman's earring and details of the clothing of both subjects have been omitted by the human cartoonist.

Mr Wood agreed that the computer cartoons were quite similar to his and that some of the line placements were remarkably good; however, he noted some further differences. In both subjects the computer had produced little rectangles in the rendition of the lower lids of the eyes, due to a tear causing a highlight in the original photograph. On the male subject, he thought there should be a shadow under the lower lip, smile lines at the side of the eyes and a small bifurcation in the line at the right of the mouth. In the cartoon of the female subject, the computer had erroneously run together the long and short lines to the left of the mouth. He felt that his own knowledge of human anatomy helped him in making these judgements.

We have not attempted to produce computer-generated caricatures. The variations in line which produce caricature have been discussed by Gombrich and Hochberg[20, 21].

Low-level processing in human vision

If a human cartoonist and a computer place their basic feature lines in roughly the same locations, might this be an indication that low-level processing in the visual system of the human cartoonist bears some correspondence to the cartoon operation used by the computer? It is interesting to note that the one-dimensional (cross-sectional) form of a valley detector (Fig. 3.3) is similar to the difference-of-Gaussian weighting function for the retinal receptive field at ganglion-cell level, while the two-dimensional form is similar to the weighting function of a cell in the visual cortex, where orientation becomes important[22, 23]. We did not model our valley detector on the human eye, but derived it from considerations of light falling on matt surfaces; in view of the correspondence between the two, could it be that the eye is particularly sensitive to surfaces identified by our postulate, that is, those which fall sharply away from the line of sight? These are clearly very important surfaces in the recognition of faces.

We note, further, the interest in Marr's idea of the 'raw primal sketch' as a first processing step in the representation of shape information in the human visual system[24]. Marr's concept is that this is produced by a spatial filter with a response similar to that of Fig. 3.3, followed by a detector of 'zero-crossings'. However, in a more recent version of this process suggested by Watt and Morgan[25] the detector of zero-crossings is replaced or supplemented by a detector of centroids. This is rather similar in its effect to our peak detector. We have also noted that scale is an important parameter in producing a recognizable cartoon. If the cartoon detector window is too small, unnecessary detail creeps in; if, on the other hand, it is too large, the shapes of the features are distorted. In human vision there are known to be several spatial filter channels operating at different scales[26].

To explore these ideas further, we carried out some experiments which involved electronic inversion of the tone scale of the images.

Inverting the cartoon

It is possible to represent a cartoon as black lines on a white background or as white lines on a black background. Our earliest cartoons[4] were in fact of the latter kind, but we noticed that when we inverted them electronically, turning black to white and white to black, they looked better. This is illustrated in Figs. 3.9–3.11; Fig. 3.9 shows an original photograph, Fig. 3.10 a black-on-white computer-generated cartoon derived from this photograph and Fig. 3.11 the white-on-black version.

Although there is exactly the same information in Figs. 3.10 and 3.11, Fig. 3.10 somehow looks more natural. Why is this? The result can be explained if

Fig. 3.9. Original photograph.

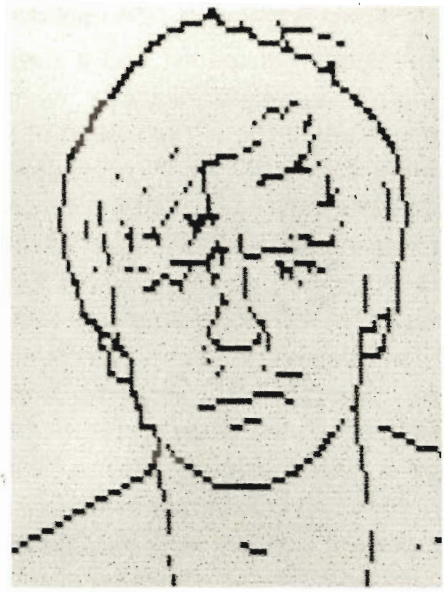


Fig. 3.10. Black-on-white computer-generated cartoon derived from Fig. 3.9.

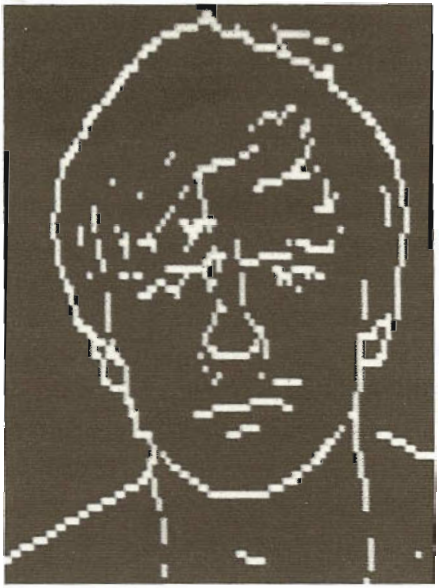


Fig. 3.11. Inversion of Fig. 3.10.

Fig. 3.12. Computer-generated cartoon using Fig. 3.11 as input.

we suppose that the early stages of human vision involve an operation similar to that of the cartoon operator, which has its primary sensitivity to valleys[27]. In a cartoon consisting of black lines on a white background, the luminance valleys in the original image are reproduced as luminance valleys in the cartoon; the depths of the valleys are not the same in the original and cartoon but their location is the same. The white-on-black inversion of this cartoon has its valleys turned into ridges. Such ridges are pointers to a different three-dimensional shape from that of a face, but by virtue of its flexibility and powers of high-level processing the human visual system is still capable of interpreting the cartoon as a face.

A black-on-white cartoon can be viewed as an image in which all the information except the valleys has been thrown away. But these valleys, according to our postulate, are very significant; they locate the bounding surfaces of smooth objects as they turn away from the line of sight. They are fundamental in gauging the size and shape of such objects.

Cartoons of cartoons

It is possible to take this experiment a step further. If the early stages of vision involve a mechanism like the cartoon operator, we can see what the output of this stage looks like by putting an image through the computer-cartooning process twice, that is by taking a cartoon of a cartoon. We can compare this double-processed image with the single-processed image or cartoon; this may give us some insight into differences between the way the eye processes a cartoon and the way it processes an ordinary image.

We took the black-on-white cartoon in Fig. 3.10 and used it as an input image for the cartoon detector. We found that it passed through the detector *unchanged*, that is the output was exactly the same as Fig. 3.10. On the other hand, when we took the white-on-black cartoon in Fig. 3.11 and passed that through the cartoon extractor, the different result shown in Fig. 3.12 was obtained. In a system using the valley detector which we have described, a black-on-white cartoon of an object thus has the interesting double property of passing through the system unaltered and of producing the same output as the original object from which it was derived. So the aim of a cartoonist may be to create a line drawing which produces the same or a similar early visual response to that of the object which it purports to represent.

Inverting the source image

If the human face-recognition system had its primary sensitivity to edges rather than to valleys, then a photographic negative should produce the same response in the early stages as the positive from which it is derived.

Inverting an image converts rising edges into falling edges (Fig. 3.1) or, equivalently, rotates the edge by 180°. A competent edge detector should find the edge location unchanged. On the other hand, if the human face-recognition system has its primary sensitivity to valleys, inverting the image will produce a different response to negatives from that to positives since the valleys are converted to ridges[28].

In Fig. 3.13 the original image of Fig. 3.9 has been inverted electronically to produce a negative. Observation indicates that it is more difficult to recognize the person concerned from the negative (though it has been suggested to us that this might be different with training).

In Fig. 3.14, we show the result of applying our cartoon operator to the inverted image of Fig. 3.13. Whereas the outlines of the head and shoulders are retained, the facial features are distorted. If we again suppose that the output of the cartooning operator is similar to the output of the early stage of the human mechanism for recognizing faces, then this output can be seen to be quite different for positive and negative images, with the positive input producing the recognizable cartoon.

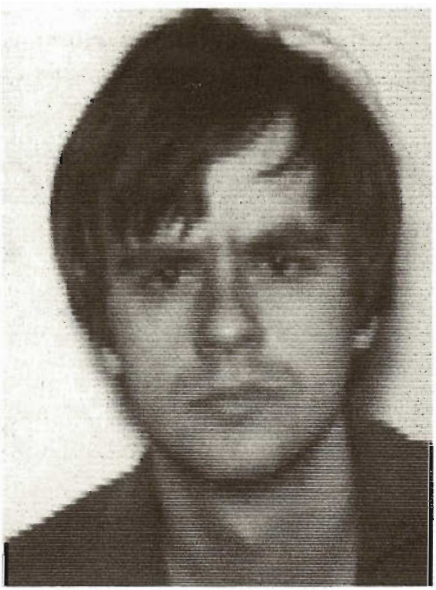


Fig. 3.13. Inversion of Fig. 3.9.

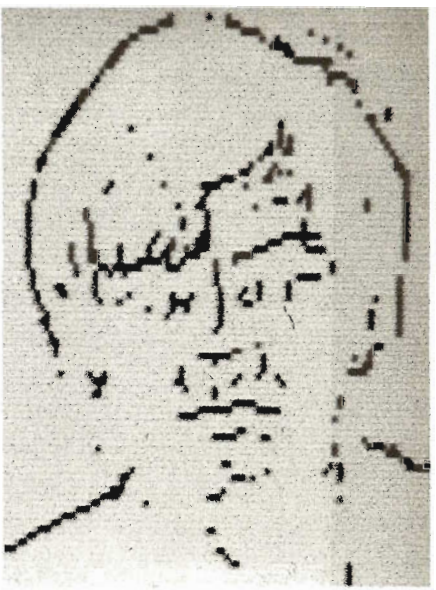


Fig. 3.14. Computer-generated cartoon using Fig. 3.13 as input.

Discussion and conclusions

We have developed a theory and ways for generating cartoons by computer. The theory states that cartoon lines should be placed wherever straight lines drawn from the television camera (or the eye of an observer) graze the surface of the human form. Analysis of the light falling on surfaces approximating skin have indicated that at such surfaces the luminance variation in the image has the shape of a valley when all local surfaces have the same luminance factor, for example at the side of the nose or at the edge of a finger held up in front of the palm of the hand. At other times, as at the skin-hair boundary or when a hand is held in front of a darker or lighter piece of clothing, a step edge of luminance occurs. What we have discovered experimentally is that a detector with a primary response to valleys and a lesser response to step edges is effective in delineating the main cartoon lines of faces and hands.

The cartoon operator which we have implemented has two component parts, the first an inverted 'Mexican-hat' filter with the required sensitivity to valleys and edges, and the second a peak detector. In practice the two are merged in the software algorithm, which has to have consideration for both speed and discrimination against noise. The working system we have produced can process still or moving grey-level images to produce the cartoons.

Using the computer-cartoonist, we have attempted to explore the nature of cartoons and the response which they set up in the human visual system. We found that when a human cartoonist was asked to draw cartoons in the style of the computer (as few lines as possible, with black fill-in), but without having seen the computer's efforts in advance, the results were remarkably similar to those of the computer. It is possible to fault the computer on some details, including its apparent knowledge of human anatomy (which was zero).

We have suggested that the degree of correspondence between the computer and the human cartoonist might be explicable if low-level processing in the human face-recognition system was similar to that used by our computer. We noted that both the filter and detection components of our cartoon extractor were similar to those being discussed in current theories of vision though they were derived from fundamental considerations of light falling on smooth surfaces and not as a model of vision. If our suggestion is correct, it provides an operational explanation for the type of early visual processing which is found in the human visual system, namely that it is good at picking out important surfaces of faces and hands in three-dimensional object space. These surfaces are ones which fall away sharply from the line of sight and

which therefore either bound the object itself, or some protrusion or indentation on it. Such surfaces are important, as has been noted, for facial recognition. Their identification may also be useful for grasping; the surfaces of fingers or hands which are detected by the cartoon operator are those which would frequently be grasped by the fingers of another person, as in shaking hands.

Since our cartoon extractor has its primary sensitivity to valleys, we noted that a cartoon drawn with black lines on a white background can reproduce these valleys at the same spatial locations as the original; it is an equivalent image in so far as the cartoon-generating system is concerned. Its equivalence has been demonstrated by taking a cartoon of the cartoon; this process leaves it unchanged. We speculated that a black-on-white line drawing which is judged to be a good likeness of something may set up the same early response in the human visual system as the thing itself.

The idea of certain images being able to pass through an image-processing system unchanged (apart from a gain constant) recalls to mind the concept of the eigenvector. Strictly speaking, this is applicable in signal theory to linear systems only, but for such systems, at any rate, there is an established notion of a signal or vector which is only amplified or attenuated and not otherwise altered by passage through the system. A sine wave, for example, is an eigenvector of a linear amplifying system, since it always emerges as a sine wave. Extending this idea rather loosely to the non-linear cartooning system we have been discussing, a black-on-white cartoon could be said to be an *eigenimage* for this system. Eigenimages form a small subset of the totality of possible images, but they include many different black-on-white line drawings, some representing real objects and some not. The suggestion we have made is that a cartoon which is a good likeness of a person is an eigenimage for the early stage of the human face-recognition system, having the further property that it approximates or corresponds with the response produced by the actual, three-dimensional, person.