Bericht 27

Formation of an Object Concept by Analysis of Systematic Time Variations in the Optically Perceptible Environment

H.-H. Nagel

IfI-HH-B-27/76

Juli 1976

Abstract

An algorithm to isolate a representation for a moving object of unknown form and size from a sequence of TV frames is described, together with results of its application to real world scenes. It is based on the assumption that a group of connected regions being jointly displaced in a systematic way from frame to frame without changing their relative position represents the image of a moving object. These assumptions relate to properties which are to a large extent independent from specific scenes and thus may appear as components in a general object concept. Implications of this approach for the conception of scene analyzing systems that "learn from observation" are discussed.

1. Introduction

Considerable experience has been accumulated about how to segment the digitized image of a static scene, e.g. a single TV frame, into areas that correspond to objects as perceived by a human observer [1 - 13]. Up to now attempts have dominated to establish the correspondence between certain segments of a scene and human notions about them by

- incorporating the required knowledge directly into the analyzing algorithm,
- interaction with a human operator,
- accessing a properly initialized data base.

Zucker, Rosenfeld and Davis [14] suggested to differentiate the knowledge required for analysis of a scene into a general, scene independent part denoted as 'general purpose models' and a specific, scene dependent part. If the latter is structurally separated into a database, changing to another scene will then only necessitate substitution of the scene dependent knowledge base - and the appropriate choice may even be guided by the information extracted with the help of scene independent knowledge. Zucker et al. proceed to differentiate their general purpose models further into (quotation from [14]) "models for general classes of local features (edges, lines, angles, etc.) that occur in many different types of scenes, as well as models that describe how such features can be grouped together into aggregates. (These aggregates may in fact not correspond to objects but they can serve as useful first guesses to guide later steps in the analysis)." The results reported by Marr [16] fit nicely into an emerging framework : "It is concluded that most of the (i.e. figure from ground) separation can be carried out using techniques that do not depend upon the particular image in question. Therefore, figure-ground separation can normally precede the description of the shape of the extracted form. Up to this point, higher level knowledge and purpose are brought to bear on only a very few of the decisions taken during processing. This relegates the widespread use of downward-flowing information to a later stage than is found in current machine vision programs, and implies that such knowledge should influence the control of, rather than interfering with, the actual data-processing that is taking place lower-down." (quotation from Marr [16]).

One may now ask : how can the scene dependent knowledge be acquired that is still necessary to evaluate abstractions into which a single digitized image of a scene has been condensed by (almost) scene independent methods ? May the analysis of e.g. a TV frame sequence which contains the development of a scene with time yield (part of) the knowledge required to bridge the gap' between the above-mentioned abstractions from a single scene image and a system-internal representation that corresponds to an object as perceived by a human observer of the original scene ? One recognizes how important it is for such an approach to condense a single digitized image into suitable abstractions without requiring already for this step the knowledge that should be extracted by analyzing a sequence of such single image abstractions. The results quoted [14,16] are taken as

an indication that the experiment alluded to above may be worthwile.

This contribution describes an attempt to separate areas corresponding to an object in motion from a static background in a sequence of TV frames solely due to the systematic changes of certain image segments from frame to frame, without any prior knowledge about specific features of the moving object. Under certain conditions - discussed later in section three - the description of a group of areas separated from the background is considered by the analyzing system as an internal representation of an object. The trajectory of this object representation relative to those parts of the scene image recognized as static background is also extracted and appended to the object representation.

The system thus arrives at a description of the relation between the moving object and the static background. Admitting more than one moving object, the relations between them could be followed in addition to those between each individual moving object and the static background. It should thus become possible to describe the development of a scene with time. The system required for this goal would build a data base from input TV frames already analyzed and use this data base for the analysis of subsequent TV frames. If a moving object stopped its motion later in the frame sequence the knowledge acquired from previous motion frames might support a static analysis e.g. along the lines reported by Marr [16]. Such a system would essentially "learn by observation". Our group started some time ago to work toward this long-term goal [17,18].

The literature about scene analysis for a sequence of scene images is much sparser than that for analyzing a single scene. Considerable work has been done on the automated tracking of cloud motion in satellite data in order to obtain reliable estimates of wind velocity [19-25]. Since the methods applied are specialized towards this goal the approaches reported cannot easily be generalized to image sequences from other scenes. The thesis of Badler [26] proposes conceptual descriptions of object movements for temporal scene analysis. Both the work of Aggarwal and Duda [22] and the thesis of Badler [26] are concerned with models of a scene sequence. More general considerations concerning the description of scene variations over space and time are presented by Uhr [27]. Wolferts [28] used sequences of frames from an airborne film camera to analyze vehicle movements in street traffic. For this purpose an interactive selection of a vehicle image in one frame and its crosscorrelation as template with a suitable section from the subsequent frame yields the shift in the position of the car image. A window around this new position is used as a template to repeat this process with the next frame in the sequence. The resulting sequence of car positions is then evaluated. Chien [29] reports a system that follows images of moving cars in TV frame sequences from a street observation by tracking prominent grey scale features in real-time. Both approaches [28,29] do not attempt to build any internal description of the scene that might enable these systems e.g. to delete the images of moving objects from TV frames and to

restore a static scene for presentation - or to present solely the moving object against a homogeneous background supplied by the system. Potter [30] describes an approach to analyze frame sequences by determining the velocity of reference edges and grouping edge points - together with those points enclosed by such edges - according to equal velocity. To what extent the method of Potter works on complicated real world images cannot be judged from the results given by him but the method seems to be susceptible to noise.

The following section describes an algorithm to identify and extract a moving object from a sequence of TV frames. The basic attitude behind this algorithm is to extract the object as an entity rather than to extract constituent parts and assemble them into an object after extraction. In addition this second section presents results obtained by applying the algorithm to sequences of real world scenes. The third section discusses in detail the assumptions underlying this algorithm in order to determine how far this approach might carry one towards the long-term goal sketched above.

2. Identifying and Isolating a Single Moving Object

2.0 Input Data

The algorithm for extracting an object representation analyses a sequence of consecutive TV frames which represent the development in time of a scene containing an object in motion. The sequence should cover a time period long enough for the moving object to be displaced in its direction of motion by at least its own extension along this direction. Then such a sequence will contain two subsequences A and B

- with a constant frame number difference between corresponding frames from each subsequence
- and the property that the background covered by the image of the moving object in each frame A_i of subsequence A is completely uncovered in the corresponding frame B_i of subsequence B or vice versa .

Figure 1 presents two frames (90 and 106) from a sequence obtained by recording a street traffic scene in real-time with a commercial monitoring vidicon camera on an AMPEX analog TV-disk. 26 consecutive TV frames have been selected from this analog TV-disk, digitized with 8-bit grey value resolution into 512 pixels per TV-line and transferred to a DECSystem-10. More details about this setup are given in [18]. Figure 2 shows a section from frame 106. This section - the position of which remained constant for all frames from this sequence - has been manually selected. It comprises 6144 pixels with 8-bit grey values in 48 lines at 128 pixels each - taking every other TV-line (from one halfpicture) on a section of 96 lines and 128 pixels which allows easy geometry-preserving enlargements without necessity to interpolate grey values. The presentation of the extraction algorithm will be exemplified by reference to the section of frames 90 through 96 as subsequence A and frames 106 through 112 as subsequence B.

2.1 Segmentation

The selected section is segmented for every frame in each subsequence using a modified Yakimovsky algorithm as described in [13]. There are only two minor differences with respect to the description in [13], the first being that the full 8-bit grey value resolution is taken into account during segmentation. The edges between the resulting regions for the section from frames A_i=90 and B_i=106 are given in figure 3. Secondly, all cracks and inner boundaries - i.e. boundaries between regions merged together [13] - are suppressed when the representation of the segmentation results are output onto disk as intermediate storage between this segmentation step and the following match step.

2.2 Region Match

The basic idea of this step consists in matching a (group of) region(s) from the segmentation of frame A_i against a (group of) region(s) from the segmentation of frame

B_i. This region match should fail in areas which belong to the background in frame A_i and to the moving object in frame B_i or vice versa.

The matching step proceeds in three substeps.
.1) The regionheaders and chain-encoded interregion boundary descriptions for the segmentation of the frame section are read back from disk for both frame A_i and the corresponding frame B_i. Then it is determined for each raster square in the frame section to what pair of regions from the segmentation of frame A_i and B_i this raster square belongs. In addition the original grey value digitizing for each pixel in the section is read in for frame A_i and the corresponding frame B_i.
.2) For each pair of regions with an overlap of more than one pixel established in the preceding substep the mean μ_A and μ_B as well as the variance σ_A^2 and σ_B^2 is determined in the overlap area on the basis of the original grey values (the index A and B refer to data from frames A_i and B_i, respectively). If the expression

$$\frac{\left(\frac{\sigma_A^2 + \sigma_B^2}{2} + \left(\frac{\mu_A - \mu_B}{2} \right)^2 \right)^2}{\sigma_A^2 * \sigma_B^2} \qquad (1)$$

is smaller than the matching threshold Tm these two regions from frame A_i and B_i are considered "match-candidates".
.3) After all match-candidates have been established, region matches are selected from among the candidates based on the following two criteria:
.3.1) If a region from frame A_i and a region from frame B_i appear in exactly one match-candidate they are accepted as a definite region-match.
.3.2) If one or both of the regions from frame A_i and B_i appear in more than one match-candidate, at least one of these regions must contribute at least half of its area to a match-candidate to accept that match-candidate as a definite region-match.

The criterion 2.2.3.2 restrains the growth for chain-matches of the following kind: region a_1 from frame A_i matches to part of region b_1 from frame B_i. (Part of) the remaining part of region b_1 matches to part of region a_2 from frame A_i. The remaining part of region a_2 matches part of region b_2 from frame B_i and so on. Figure 4 shows the largest group of chain-matched regions when working with frame A_i=90 and B_i=106.

2.3 Object-Candidate Selection

Figure 5 presents all regions from frame B_i=106 not matched to a region in frame A_i=90. It is obvious from inspection of figure 5 that the set of all unmatched regions from one frame section may not be a good approximation to a representation of a moving object but that some subset could serve this purpose as an "object-candidate". It can be seen that selecting the largest 4-connected group of unmatched

regions may serve as working criterion for isolating this object-candidate. Figure 6a represents the subset of unmatched regions selected by this criterion from figure 5.

2.4 Accumulating Evidence for a Moving Object

The procedure described in steps 2.2 and 2.3 can be performed for frames A_i and B_i as well as for frames A_{i+1} and B_{i+1}. This will yield two object-candidates, one from frame B_i and one from frame B_{i+1} (see figure 6b for the result from B_{i+1}=107). If both object-candidates do indeed represent the same moving object one expects to be able to transform them into each other taking the appropriate object motion into account. The simplest object motion in this context is a size and form preserving translation in a plane perpendicular to the optical axis of the TV-camera. At this stage of analysis, the translation of the object images from frame B_i to frame B_{i+1} is unknown with respect both to direction and magnitude. The boundaries of all regions in the object-candidate from frame B_i are taken as a template and crosscorrelated to the boundaries of all regions in the object-candidate from frame B_{i+1}. The region boundaries rather than the grey values of the object-candidates have been chosen for the crosscorrelation due to two observations:

i) The crosscorrelation of a (binary) boundary/nonboundary representation can be performed faster than the crosscorrelation of grey values.
ii) With respect to alignment, boundaries are more prominent features than regions with inhomogeneous grey values and should thus result in a more prominent crosscorrelation peak.

Wolferts [28] makes a similar observation for the crosscorrelation maximum of the spatial derivatives of two digitized grey value distributions. Figure 7 represents the region boundaries for the object-candidates of figure 6 and figure 8 the resulting crosscorrelation. The maximum of this crosscorrelation corresponds to an object-candidate translation of five rasterpoints along the horizontal axis from frame B_i=106 to frame B_{i+1}=107.

One could assume that the difference between the center of gravity of object-candidates in frame B_i and B_{i+1} would yield the same value at much less computational cost. Since the form of object-candidates may vary somewhat from frame to frame due to statistical fluctuations in the size and form of regions, the positional shift of the center of gravity of object-candidates does not appear to be as reliable as the positional shift of the maximum for the boundary crosscorrelation. Figure 9 represents the outer boundaries for object-candidates extracted from frame 106 through 112 with the center of gravity marked for each object-candidate. One can clearly see in figure 10 that the shift of the boundary crosscorrelation maximum is much smoother than the shift of the center of gravity.

2.5 Extraction of an Object Representation

Having accumulated sufficient evidence for a moving object, a representation for it can be extracted by superimposing the different object-candidates, taking into account their relative position and then thresholding the resulting superposition. Figure 11 represents how often each raster square appears in this superposition if six object-candidates are translated back to the position of the one from frame B_i=107. If it is requested that only those raster squares belong to the object form representation which appear in at least the fraction p of all superimposed object-candidates the result of figure 12 is obtained for p = 2/3 . The dark side windows have been matched sufficiently often to the background street and thus are not considered part of the object. The wide missing part of the car body is due to a long region in this part of the car image - see e.g. figure 3b - which corresponds to the transition from the light car body to the dark shadow below the car. This transition region has a grey value roughly equivalent to the background street. It is sometimes matched to this background and therefore does not belong sufficiently often to an object-candidate.

This object form representation together with the knowledge about the object-candidate positions can be used to extract the grey values of the object image from the different frames of subsequence B and to average them. The resulting grey value representation for this object image is shown in figure 13.

To prove that this approach is not tailored to the scene used so far as an example, a new scene has been taken with a high quality plumbicon camera. Figure 14 shows two frames of a sequence where a young lady walks in front of a blackboard. Figure 15 shows a section of the same size as in figure 2. Applying the segmentation process with the same threshold and region-merging parameters as for the car sequence and matching the resulting regions with the same matching threshold parameter yields the object-candidates of figure 16. Superimposing these object-candidates with their proper relative translation and thresholding the resulting distribution with p = 2/3 yields figure 17. From this the grey level representation of figure 18 is obtained.

2.6 Operation Data

The program has been written in PASCAL for a DECSystem-10 with a maximum of 68 Kwords at 36 bits available for a user. The code including constants required 9 Kwords, the global variables and working spaces 19 Kwords, heap for segmentation and result descriptions 38 Kwords, and I/O buffers 2 Kwords. With this core allotment two subsequences with up to six frame sections each could be handled. The segmentation step required between 90 and 150 seconds. The match step including the selection of an object-candidate

required about 90 seconds and the crosscorrelation in the Evidence Accumulation step about 4 minutes. These times include I/O and refer to a KA-10 processor with no other timesharing user active but with an abnormal, not yet completely understood system overhead of about 35-40 % which is tentatively attributed to the fact that this program required all of user core and some system jobs are intermittently activated.

3. Assumptions Underlying the Described Algorithm

The presentation of the preceding section either explicitly mentions or implies a number of assumptions which will now be discussed regarding the following considerations:
- To what extend does an assumption which has been made to obtain the results presented restrict the applicability of the described approach?
- Is it possible to drop or weaken some of these assumptions in order to make the approach more generally applicable?

3.1 The Notion of an 'Object'

The algorithm presupposes the existence of a
(1) single entity of
(2) reasonable size
(3) moving smoothly and continuously
(4) in front of
(5) a sufficiently contrasting background.

The first three of these assumptions refer to context independent object properties whereas the latter two refer to the relation between object and background, i.e. are somewhat context dependent.

3.1.1 Without any specific knowledge about what may be contained in the scene to be analyzed it is reasonable to assume that a group of contiguous areas in the image sequence which are displaced without changing their relative position represents "one object". Observation of a longer sequence - i.e. development of a scene over a longer period - may well lead to the further conclusion that such an object has to be differentiated into two or more constituent objects. This conclusion implies that the system described here has been embedded into a more powerful enclosing system with the capability to recognize and treat "compound objects" consisting of several constituent objects according to the object concept defined by 3.1.1 through 3.1.5. If the image contrast between object and background is low the extraction of an object form representation might require analysis of more frame sections than necessary with higher contrast. Therefore, the length of a frame sequence during which a group of contiguous areas must be considered to represent a single object-candidate depends also upon the image contrast between object and background. If several moving objects have already been identified separately in a longer frame sequence one could extrapolate their trajectories - assuming 3.1.3 is valid - and exclude sections from analysis where the images of two or more objects occlude (parts of) each other.

3.1.2 The assumption of a single object in a frame section implies that the object is small enough compared with the size of the entire frame to allow the selection of a suitable section around the object. The size of the frame section chosen in the described implementation has been determined by several considerations. Dominant was the desire to work with raw TV digitizings - not averaged in any

down-scaling procedure - in order to be fully exposed to the intrinsic variations of the data. The additional requirement to keep the full description of segmentation results from at least four frame sections together with a debugging version of the match program in core then roughly determined the acceptable number of pixels in the section.

3.1.3 The algorithm assumes that the object image is displaced from frame to frame by only a few raster squares. This influences the size of the frame section and the search range during the crosscorrelation. If one has access to a halfpicture every 20 msec this assumption is clearly justified for a large set of scenes. Taking the center of gravity displacement of subsequent object-candidates as a first guess one might reduce the search range for the crosscorrelation in the evidence accumulation step. This search range is currently set to ten raster squares in both directions for each coordinate.

3.1.4 The basic attitude behind this algorithm is to extract the object as an entity rather than to extract constituent parts and assemble them into an object after extraction. It is obvious that the "extraction as entity" approach will fail if the object is partly occluded by static components of the scene or some other moving object. Again, if the scene can be observed long enough, one may be able to extract the object representation in a suitable scene section and extrapolate its trajectory back into or through the difficult part.

3.1.5 For the purpose of this investigation, occlusion is equivalent to vanishing contrast between (parts of) the moving object and (parts of) the surrounding static background or some other moving object. Unless the object can be observed against a different background, even a human observer might have difficulties in such situations.

3.2 Selection of a Frame Section

Assuming that the moving object occupies only a small fraction of the field of view relates to another problem, namely to focus attention to that section within the frame 'where the action is' . We have developed an algorithm [31] to identify the section of a TV frame where a moving object is suspected. The absolute difference in grey value of pixels
from two neighboring frames - averaged over two adjacent raster squares in a line - is calculated for all TV-lines and the distribution of such differences is determined as averaged over all TV-lines in a frame. Comparison of this averaged distribution with the one obtained for a small group of adjacent TV-lines will show characteristic deviations if this group of adjacent TV-lines covers (part of) a moving object. This algorithm uses the assumption that the object does not cover a large fraction of the field of view and in addition that not too many smaller objects are present in the field of view. Otherwise the grey value difference distribution averaged over the entire frame may not be representative for the static background alone. If these

assumptions are justified the problem of more than one moving object in a frame sequence may be reduced to the single object case as long as these objects do not occlude (parts of) each other.

3.3 Frame Number Difference

Determining a suitable frame number difference between subsequences to be compared is another problem that may be solved essentially by the same approach as sketched for selection of a frame section around a moving object. The 'significant' differences between subsequent frames that guide the section search will show systematic positional changes with frame number. It should thus be possible to determine the frame number distance between frames where something changes in a selected section during one subsequence with no changes significant with respect to a moving object in the same section from another subsequence. It should be pointed out that this frame number difference corresponds to a very important characteristic of situations with moving objects: after what time is it possible to state a significant change ?

3.4 Segmentation

One might question wether it is really necessary to segment a frame section and match the resulting regions rather than matching the section pixel by pixel to the section of a corresponding frame from the other subsequence. Provided the contrast between the object and the background is large enough, pixel by pixel subtraction and selection of the largest group of 4-connected pixels which show a grey value difference larger than a threshold will yield an object-candidate of acceptable quality as can be judged from figure 19. Preliminary experience tends to support the hypothesis that a region match of segmented sections is a more robust approach than the considerably simpler procedure sketched above. This is in accordance with results reported by Price and Reddy [32]. Especially when the frame section contains textured areas the approach based on segmentation may still work whereas the simple subtraction is expected to give trouble. It has to be investigated whether the simple approach can be used to extract object-representations when the contrast is favourable, restricting the algorithm described in section 2 of this contribution to the initial isolation of object-candidates and to frame sections with low contrast.

3.5 Extraction : Comparing Object with Background

Figure 19 suggests that the isolation of an object-candidate might be rather simple - which it indeed may be if appropriate frames are compared. It might be worthwile to recapitulate some of the experiments which eventually led to the algorithm described in section 2.2 and 2.3. Originally I attempted to match regions between sections from neighboring frames with the intention to detect systematic changes in the position of some subgroup of matched regions.

It then seemed easier to identify those regions which did not shift from frame to frame, i.e. which correspond to the static part of the scene, by superimposing region boundaries from segmentation of a sequence of frame sections. Discussions about intermediate results fom this attempt [33] led to the suggestion to average the raw digitizings of a frame sequence. In pursueing this approach I averaged all 26 digitized frames, segmented the section from the averaged picture and attempted to match the resulting regions against those from the section of a single raw digitizing frame.Larger groups of regions which belong to the image of the moving object in the single frame were indeed not matched to regions of the averaged frame section. Closer investigations showed, however, that the part of the moving object which is not significantly structured along the direction of motion - as e.g. the car body below the side windows in figure 2 - influenced the averaged picture sufficiently to allow a match with corresponding regions from the car. Moreover, questions had to be answered regarding where to begin averaging and over how many frames one should average. From these investigations resulted the idea to compare a frame with a single previous - or later - one where the moving object can be directly contrasted against the static background rather than against parts of itself. This approach will work if the static background is sufficiently homogeneous and of sufficient contrast compared to the moving object. To find quantitative measures for "sufficient" is subject of a further investigation. It can be guessed from comparison between figure 19a and 19b that a grey value difference of 10 (at a total grey value range from 0 to 255) seems to be sufficient even if the difference is exceeded frequently at isolated pixels due to noise and lineJitter in the digitizings electronics.

3.6 Matching Threshold

The expression (1) to be compared with the matching threshold Tm is formed in analogy to the edge confidence measure given by Yakimovsky [9]. However, in contrast to the expression given by Yakimovsky, the exponentiation by the number N of pixels in the overlap area of two regions from different frames has been suppressed. This is an empirical approach taken to simplify the decision with no apparent detrimental results. To use a correct statistical measure one has to compare

$$\frac{\left(\frac{\sigma_A^2 + \sigma_B^2}{2} + \left(\frac{\mu_A - \mu_B}{2}\right)^2\right)^{2N}}{\sigma_A^{2N} * \sigma_B^{2N}} \qquad (2)$$

with a confidence threshold that is a function of N . It is assumed that this dependency upon the number N of pixels in the overlap between two regions from different frames will become relevant only in the limit of very low contrast between object and static background. This question has to be investigated yet.

3.7 Object-Candidate Selection

If the assumption of a single moving object in the frame section is justified the criterion of step 2.3 seems to be reasonable. However, even if a small number of moving objects is admitted in the same section, simple criteria like size, center of gravity position, the first few moments about the center of gravity, and average grey value should enable one to combine the most likely object-candidates for the following step. If enough objects of very similar size, color, and shape move too close to each other, a human observer will have difficulties, too, to follow each individual object.

3.8 Superposition of Object-Candidates

Here the basic assumption is that the object image does only change its position due to translation. No rotation nor changes in size or shape are allowed. It must be investigated, however, to what extent such changes of an object image can be admitted provided the frame to frame differences due to such changes remain small.

4. Conclusion

An algorithm to isolate a representation for a moving object of unknown form and size from a sequence of TV frames has been described. It is based on the assumption that a group of connected regions being jointly displaced in a systematic way from frame to frame without changing their relative position represents the image of a moving object. These assumptions relate to properties which are to a large extent independent from specific scenes and thus may appear as components in a general object concept. Using these assumptions to isolate an object representation from a sequence of frames and to determine the development of the relation between the isolated representation and static parts of the image over time let it appear possible to conceive a system which will acquire some of the knowledge about a scene which had to be supplied hitherto explicitly by the system designer or operator.

Under conditions which can be verified by the analyzing system this approach promises to be applicable to a wide range of interesting scenes. This algorithm could be applied not only to digitizings of TV frames but of other frame sequences, too, provided the changes in the object image from frame to frame are small enough to justify the assumption of a continuous object motion.

It requires the following parameters:
for the segmentation [13]
- two reference coordinate pairs selecting the section
- the edge confidence threshold
- the region dissimilarity threshold

and for the isolation of an object representation
- the frame number difference between frames to be compared
- a matching threshold for the region match
- a threshold fraction to determine the object form from superimposed object-candidates.

It has to be explored in detail how reliable the determination of the reference coordinate pairs for the frame section and of the frame number difference parameter will be. Since the computation required for this method can easily be performed in a minicomputer, selection of frame sections for more detailed analysis can be delegated to the digitizing process being executed on a minicomputer. If feedback from the isolation of an object-candidate to the digitizing phase is possible, selecting a frame section of interest can be simplified by extrapolating the trajectory of an object-candidate already identified - which is admitted under the continuity assumption 3.1.3 - . This is especially attractive if the TV frame sequence has been buffered on an analog TV-disk and the digitizing electronic comprises an electronic windowing feature with programmable window limits as in our case.

If the segmentation step can be avoided, everything up to and including the object-candidate selection could easily be delegated to a minicomputer. It remains to be investigated under what conditions the segmentation step can be omitted. It will be required, however, when the "extraction as entity" approach fails and the match of partial object descriptions in situations with occluding obstacles or vanishing background contrast has to be attempted.

Although the examples shown are of very different form and taken by different cameras under differing illuminations, only the reference coordinates for the frame section and the frame number difference had to be adapted to yield acceptable results. This can be taken as a hint that the method is robust enough to warrant further experiments. It should be pointed out that this method may be applied to establish the correspondence between reference points for a moving object in sequences of stereo frames.

The detailed discussion of the assumptions underlying this approach indicated a number of specific questions which are currently investigated. It may well be that this specific incentive for further research is one of the merits of the approach reported.

5. Acknowledgements

Continuous efforts by all members of our group - including R. Bertelsmeier, P. Cord, I. Heer, H. Kemen, B. Neumann, and B. Radig - have been necessary to maintain and improve our experimental apparatus without which this investigation would have been significantly more cumbersome and less exciting - if possible at all. I gratefully acknowledge a number of important discussions with members of our group, especially B. Neumann and B. Radig, about the subject of this contribution. Miss L. Dreschler helped in programming during the last phase of this investigation. I am grateful to Mrs. R. Jancke for her support in preparing this contribution.

Our digital picture processing laboratory has been set up with financial support from the Federal Ministry for Research and Technology and from the Freie and Hansestadt Hamburg, both extended through the university. This is gratefully acknowledged.

References

1. A. Rosenfeld, Picture Processing by Computer, Academic Press, New York, 1969.

2. R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, A. Wiley Interscience Publ., New York, 1973.

3. A. Rosenfeld and A.C. Kak, Digital Picture Processing, Academic Press, New York 1976

4. A.J. Thomas and T.O. Binford, Information Processing Analysis of Visual Perception, Stanford Computer Science Department, Report STAN-CS-74-408, June 1974.

5. S.W. Zucker, Region Growing : Childhood and Adolescence, Computer Science Center, University of Maryland, College Park, Maryland
Technical Report TR-370, April 1975

6. H.G. Barrow and J.M. Tenenbaum, Representation and Use of Knowledge in Vision, ACM SIGART Newsletter No. 52, June 1975, p. 2.

7. Contributions under Section 9 "Visual Information Processing" and Section 10 "Robots and Productivity Technology" in the Advance Papers of the 4th International Joint Conference on Artificial Intelligence (= 4IJCAI), Tbilisi, Georgia/USSR, September 3-8, 1975.

8. J.W. McKee and J.K. Aggarwal, Finding the Edges of the Surfaces of Three-dimensional Curved Objects by Computer, Pattern Recognition 7, 1975, 25.

9. Y. Yakimovsky, Boundary and Object Detection in Real World Images, Technical Memorandum 33-709, Jet Propulsion Laboratory, Cal. Institute of Technology, Pasadena, Cal., November 15, 1974. and 4IJCAI, p. 695.

10. R.B. Ohlander, Analysis of Natural Scenes, (Ph.D. Thesis), Computer Science Department, Carnegie-Mellon-University, Pittsburgh, Pa., April 1975

11. Y. Yakimovsky, On the Problem of Embedding Picture Elements in Regions, Technical Memorandum 33-774, Jet Propulsion Laboratory, Cal. Institute of Technology, Pasadena, Cal., June 1, 1976.

12. B.J. Schacter, L.S. Davis and A. Rosenfeld, Scene Segmentation by Cluster Detection in Color Spaces, ACM SIGART Newsletter No. 58, June 1976, p 16.

13. H.-H. Nagel, Experiences with Yakimovsky's Algorithm for Boundary and Object Detection in Real World Images, Proc. 3rd International Joint Conference on Pattern Recognition, November 8-11, 1976, Coronado, California
and
Bericht Nr. 23, Institut fuer Informatik, Universitaet Hamburg, Hamburg, March 1976

14. S.W. Zucker, A. Rosenfeld and L.S. Davis, General Purpose Models - Expectations about the Unexpected, 4IJCAI, p. 716. and ACM SIGART Newsletter No. 54, October 1975, p. 7

see also

15. A. Rosenfeld, Non-Purposive Perception in Vision, NTG/GI Fachtagung "Cognitive Verfahren und Systeme", Hamburg, April 11-13, 1973, (Th. Einsele, W. Giloi, and H.-H. Nagel, eds) Lecture Notes in Economics and Mathematical Systems, vol 83, Springer Verlag, Berlin-Heidelberg-New York 1973, pp. 349-373

16. D. Marr, Early Processing of Visual Information, MIT Artificial Intelligence Memo No. 340, December 1975

17. See ACM SIGART Newsletter No. 50 of Februar 1975 containing a short notice about our work on page 3 (from H.-H. Nagel, Institut fuer Informatik der Universitaet Hamburg).

18. H.-.H. Nagel, Towards Analyzing Sequences of Scenes Containing Objects in Motion, Contribution to the 2nd meeting GI-Fachgruppe Artificial Intelligence, October 7, 1975, University of Dortmund, Berichte des Instituts fuer Informatik Dortmund (G. Veenker, ed.), Nr. 13/75, p. 126.

19. J.A. Leese, C.S. Novak and V.R. Taylor, The Determination of Cloud Motion Patterns from Geosynchronous Satellite Image Data, Pattern Recognition, vol. 2, December 1970, pp. 272-292.

20. E.A. Smith and D.R. Phillips, Automated Cloud Tracking Using Precisely Aligned Digital ATS Pictures, IEEE Trans. Computers, vol. C-21, July 1972, pp. 715-729.

21. D.J. Hall, R.M. Endlich, D.E. Wolf and A.E. Brain, Objective Methods for Registering Landmarks and Determining Cloud Motions from Satellite Data, IEEE Trans. Computers, vol. C-21, July 1972, pp. 768-776.

22. J.K. Aggarwal and R.O. Duda, Computer Analysis of Moving Polygonal Images, IEEE Trans. Computers, vol. C-24, 1975, pp 966-976

23. A.A. Arkins, R.C. Lo, and A. Rosenfeld, An Evaluation of Fourier Transform Techniques for Cloud Motion Estimation, Comp. Sci. Center, Univ. Maryland, College Park, Maryland, Technical Report TR-351, January 1975

24. R.C. Lo, The Application of a Thresholding Technique in Cloud Motion Estimation from Satellite Observations, Comp. Sci. Center, Univ. Maryland, College Park, Maryland, TR-357, February 1975

25. R. Green, G. Hughes, C. Novak, and R. Schreitz, The Automatic Extraction of Wind Estimates from VISSR Data, in Central Processing and Analysis of Geostationary Satellite Data (C.L. Bristor, ed), NOAA Technical Memorandum NESS 64, Washington, D.C., March 1975

26. N.I. Badler, Temporal Scene Analysis: Conceptual Descriptions of Object Movements, Technical Report No. 80, Department of Computer Science, University of Toronto, February 1975.

27. L. Uhr, The Description of Scenes over Time and Space, National Computer Conference, June 4-8, 1973 AFIPS vol. 42, Montvale, New Jersey, p. 509

28. K. Wolferts, Ein Interaktives Verfahren zur teilautomatischen Auswertung von Luftbildern fuer Verkehrsanalysen, Proceedings of the NTG/GI Fachtagung "Cognitive Verfahren und Systeme", Hamburg, April 11-13, 1973, (see [15]) , p. 307
and
Special Problems in Interactive Image-Processing for Traffic Analysis, 2nd International Joint Conference on Pattern Recognition, Copenhagen, August 13-15, 1974, p. 1

29. R.T. Chien and V.C. Jones, Acquisition of Moving Objects and Hand-Eye Coordination, 4IJCAI, p. 737.

30. J. Potter, Scene Segmentation by Velocity Measurements Obtained with a Cross-Shaped Template, 4IJCAI, p. 803.
and earlier reports quoted in that contribution

31. Detlev Militzer, Extraktion von Bildausschnitten mit signifikanten Aenderungen in Fernsehbildfolgen, Studienarbeit, Institut fuer Informatik der Universitaet Hamburg, September 1976

32. K. Price and R. Reddy, Change Detection in Multi-Sensor Images, Tenth International Symposium on Remote Sensing of Environment, Ann Arbor, Michigan, October 1975

33. Thomas Bonde, Untersuchungen zur Zuordnung von Bildbereichen bei segmentierten Fernsehbildsequenzen, Studienarbeit, Institut fuer Informatik der Universitaet Hamburg, September 1976

Figure Captions

Fig. 1 Frame 90 (Fig. 1a) and Frame 106 (Fig. 1b) from a sequence of video frames recorded on an AMPEX analog TV-disk. The frame sequence represents a street intersection with traffic, observed from our laboratory window by a commercial monitoring vidicon camera.

Fig. 2 Section of 96 lines, at 128 pixels each, of frame 106 enlarged to full frame format, reproduced on a TV-monitor with the help of a Thomson-Houston solid-state image storage tube from digitized data.

Fig. 3 Segmentation results of equal sections from frame 90 (= Fig. 3a) and frame 106 (= Fig. 3b) obtained with modified Yakimovsky algorithm [13].

Fig. 4 The largest set of chain-matched regions obtained in the match of the section from frame 90 to the section from frame 106. A 'commercial at' (@) is printed for every pair of matched pixels. Unmatched pixels from a region in frame 90 which is only partially matched to regions from frame 106 are printed as alphabetic characters - see e.g. part of the street in frame 90 covered by the front of the moving car in frame 106 which appears as P in the lower left. Unmatched pixels from a partially matched region of frame 106 appear as nonalphabetic characters. Note the two holes around column 72 between lines 5 and 12 which correspond to white center stripes matched to their counterpart in the other frame without connection to the enclosing darker street background.

Fig. 5 The set of all unmatched regions of frame 106 in a match attempt to frame 90. The use of a question mark to represent different regions in this figure is only due to the printing routine and does not indicate any other conclusions.

Fig. 6 Object-candidates which are defined here as the largest group of 4-connected unmatched regions from frame 106 when matched to frame 90 (fig. 6a) and from frame 107 when matched to frame 91 (Fig. 6b).

Fig. 7 Region boundaries for the object-candidates from Fig. 6a and 6b, respectively.

Fig. 8 Result of crosscorrelating the region boundaries from Fig. 7a with those from Fig. 7b. The vertical shift is denoted as Δy and varies from $\Delta y = -10$ raster squares in the top line to $\Delta y = +4$ raster squares (i.e. down) in the bottom line.

The horizontal shift is denoted as Δx and varies from $\Delta x = -10$ in the left column to $\Delta x = +10$. The topmost of the three numbers for each Δx and Δy indicates the number of overlapping horizontal edges, the middle one indicates the number of overlapping vertical edges, and the bottom one gives the sum. The maximum is obtained for $\Delta y = 0$ and $\Delta x = 5$.

Fig. 9 Object-candidates obtained for the sections from frame 106 through 112 (Fig. 9a through 9g, respectively). Only boundaries to the background are indicated. The lines are drawn between the outermost raster point of the object-candidate and the next raster point of the background. The circled cross indicates the position of the center of gravity of the object-candidate. The object-candidates are shown in their proper position relative to the constant vertical section edges.

Fig. 10 x-coordinate of object-candidate as a function of frame number for frames 106 through 112 of street intersection scene.

Fig. 11 Frequency distribution of raster squares for superposition of all object-candidates after compensating their shift from frame to frame.

Fig. 12 Object form representation, obtained from Fig. 11 by retaining only those raster squares which appear in at least 4 out of 6 object-candidates ($p = 2/3$).

Fig. 13 Object grey value representation extracted from frames 108 through 112 of street traffic scene. Note that averaging over 5 frames let appear wheels discernible which cannot be seen in a single frame enlargement as Figure 2. (The available core was just insufficient to handle six frames from 107 through 112 when the extraction of a grey value representation had been added to the program. Therefore, only five frames were used with $p = 3/5$ to produce this picture.)

Fig. 14 Frame 435 (Fig. 14a) and frame 457 (Fig. 14b) of an indoor scene recorded without artificial illumination by a Plumbicon camera on an AMPEX analog TV-disk.

Fig. 15 Section of the same size as in Fig. 2 from the digitizings of the frame shown in Fig. 14a, reproduced on a TV-monitor with the help of a Thomson-Houston solid-state image storage tube.

Fig. 16 Object-candidates extracted from the sections of frame 457 through 462 (Fig. 16a-f, respectively) by matching them to corresponding sections from frames 435 through 440.

Fig. 17 Object form representation, obtained from superposition of object-candidates shown in Fig. 16a-f after compensating their frame to frame shift and retaining only those raster squares which appear in at least 4 out of the 6 object-candidates ($p = 2/3$).

Fig. 18 Object grey value representation extracted from frames 457 through 462 of indoor scene, corresponding to object form as shown in Fig. 17. Please, note that the object representation comprises areas which are both lighter and darker as the background - although the hair region is difficult to discern in this reproduction.

Fig. 19 Raster squares with grey value differences in excess of 10 (Fig. 19a) and of 15 (Fig. 19b) between the same sections of frame 90 versus frame 106 as used in Figure 3. Grey values had been digitized with 8 bits. The car representation results from frame 106 whereas the pixel group at the left hand side is due to the car front appearing in this section of frame 90.

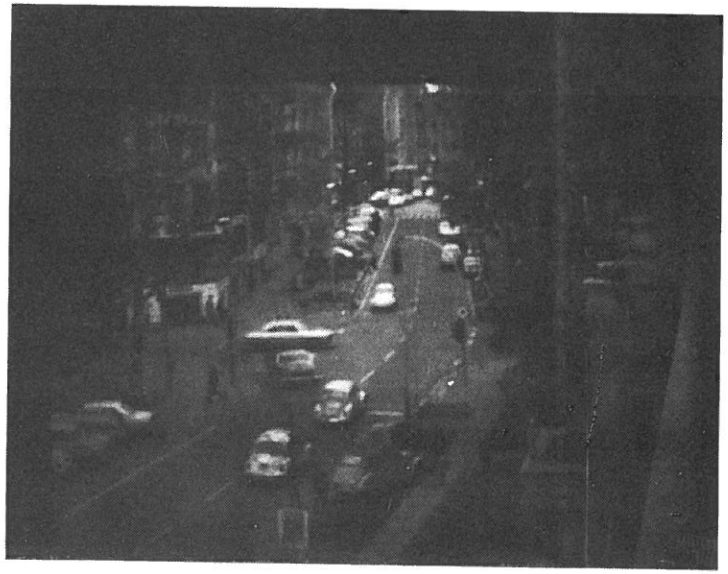

Fig. 1 a

Fig. 1 b

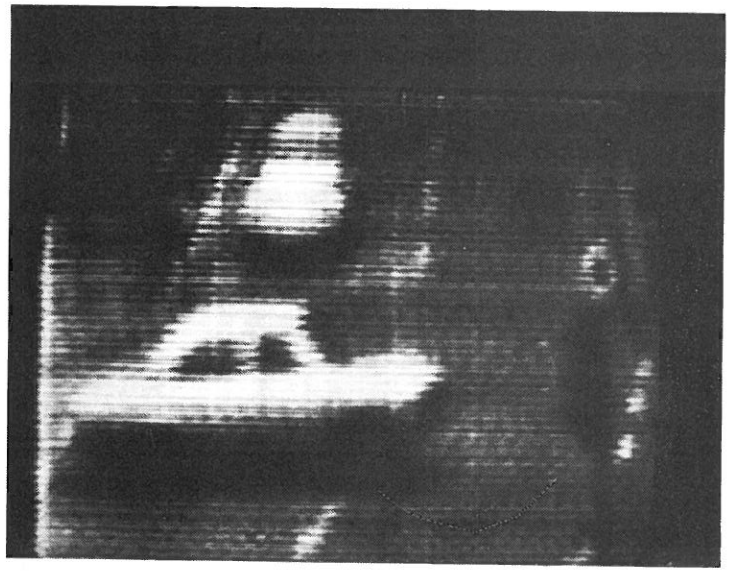

Fig. 2

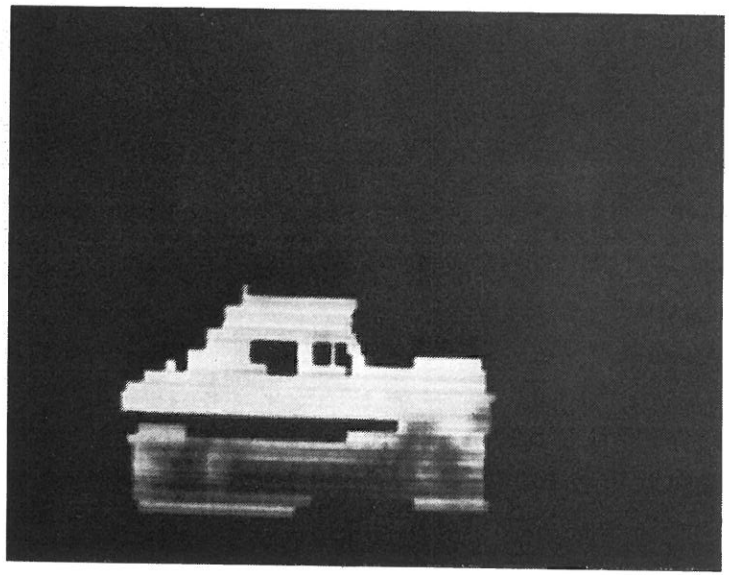

Fig. 13

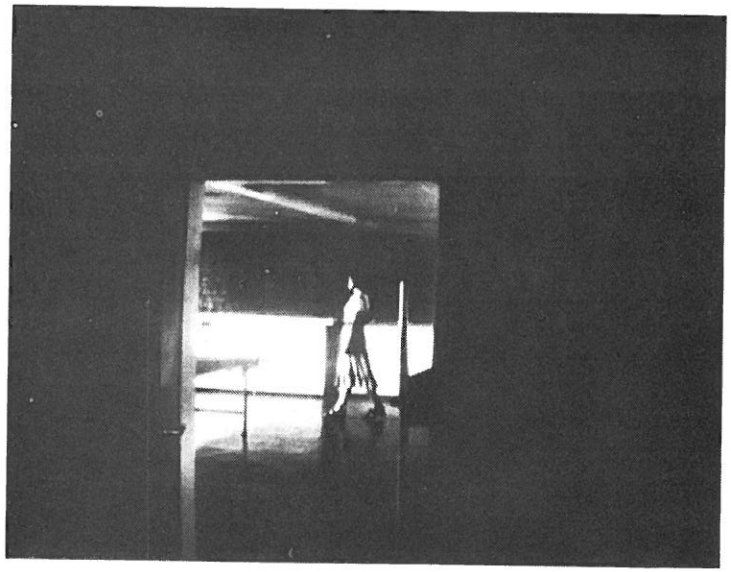

Fig. 14a

Fig. 14b

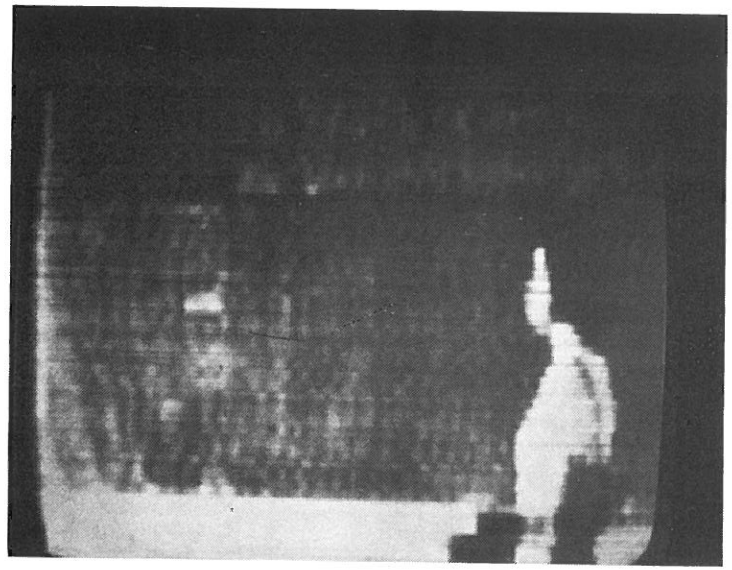

Fig. 15

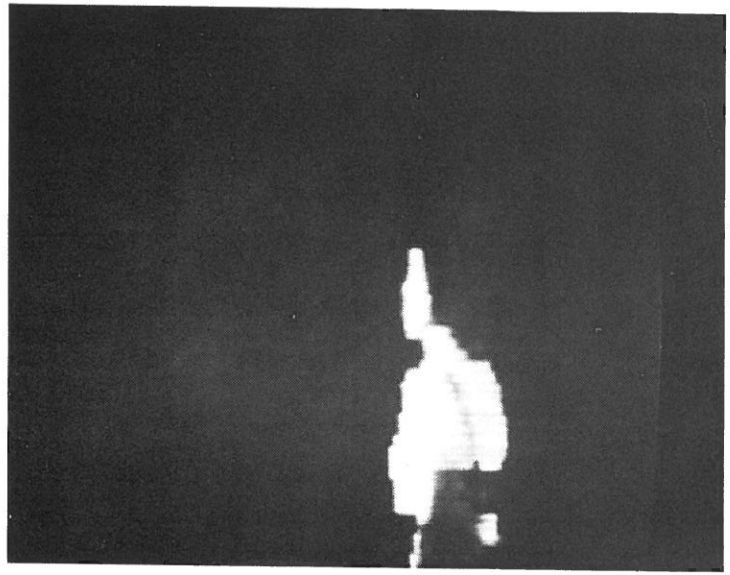

Fig. 18

VERSION V3A41 AM 11-JUN-76 UM 9: 8124,340 UHR FUER HALBBILD-AUSSCHNITT AUS DSK : BILD ,V00 [000326,000102]
ERSTE ZEILE = 232 ERSTE SPALTE = 231 MIT THRESHOLD = 100,0 UND GESCHWELLE = 2,0 AUSGABE VON GEBIETEN

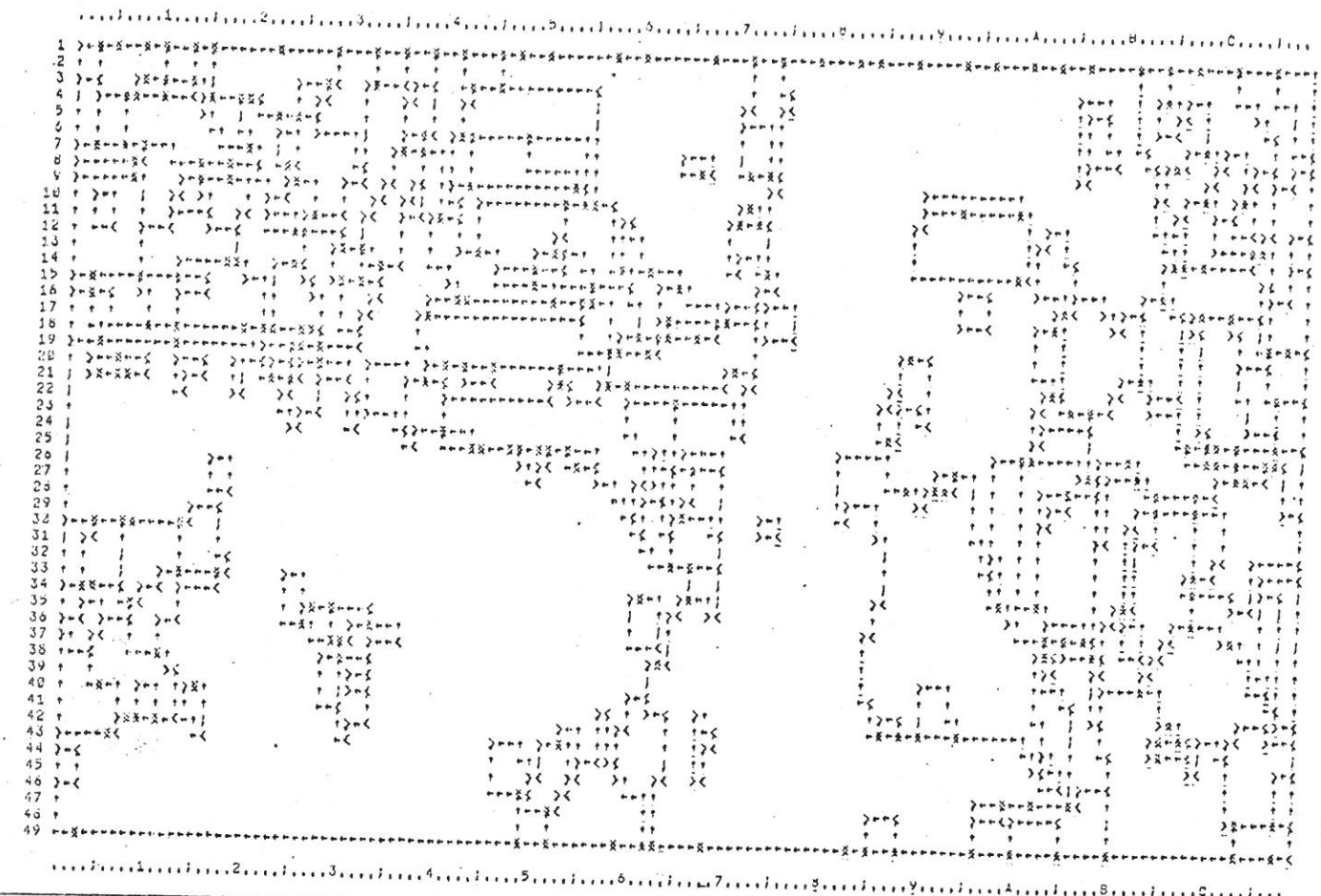

Fig. 3a

VERSION V3A41 AM 11-JUN-76 UM 19:23: 8,350 UHR FUER HALBBILD-AUSSCHNITT AUS DSK : BILD ,V00 [000326,000102]
ERSTE ZEILE = 232 ERSTE SPALTE = 231 MIT THRESHOLD = 100,0 UND GESCHWELLE = 2,0 AUSGABE VON GEBIETEN

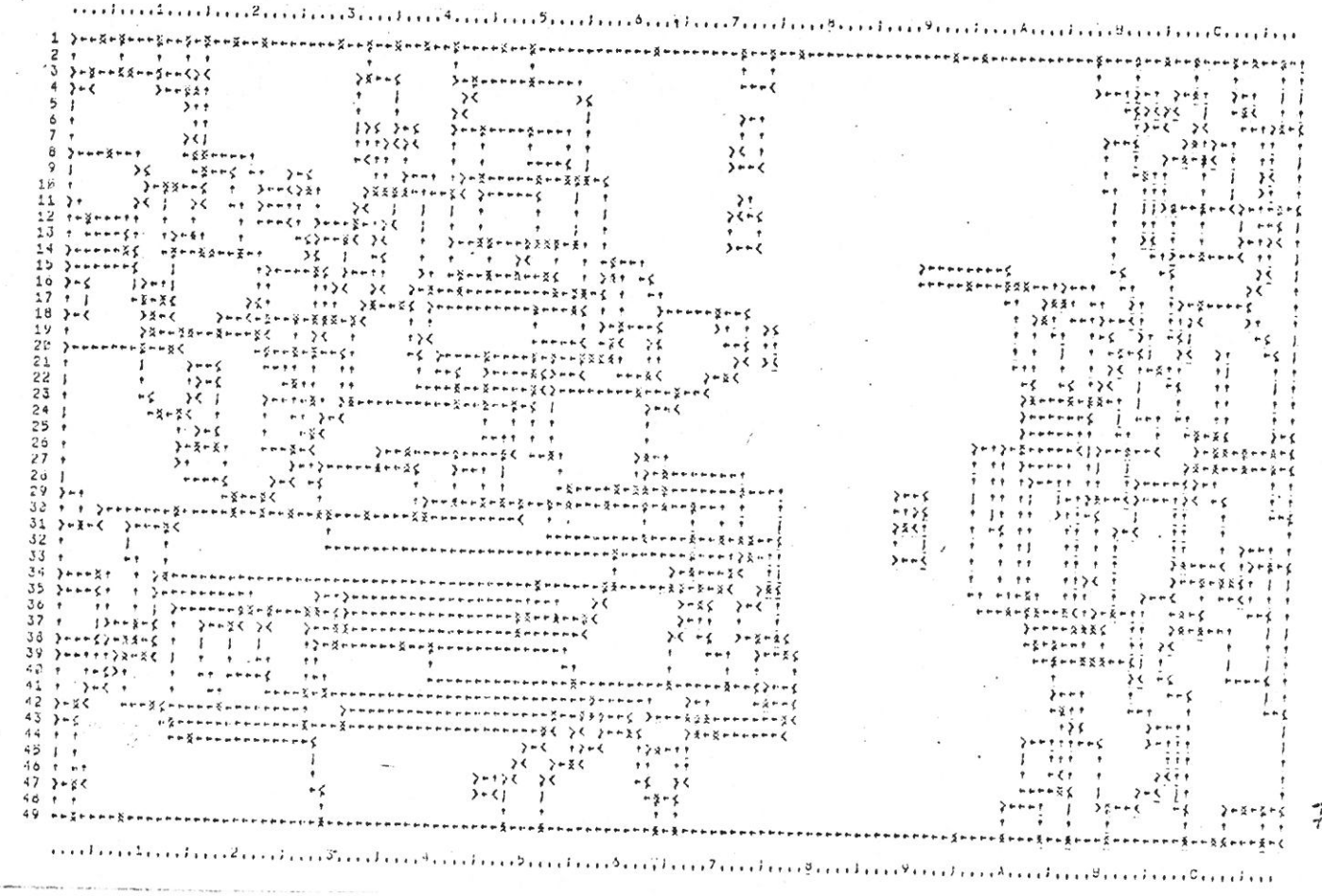

Fig. 3b

VERSION V3A41 AM 11-JUN-76 UM 9: 0:24,340 UHR FUER HALBBILD-AUSSCHNITT AUS DSK : BILD ,V90 [000300,000100]

ERSTE ZEILE = 232 ERSTE SPALTE = 231 MIT THRESHOLD = 100,0 UND G-SCHWELLE = 2,0

Fig 4

VERSION V3A41 AM 11-JUN-76 UM 19:20: 8,360 UHR FUER HALBBILD-AUSSCHNITT AUS DSK : BILD ,VD0 [000300,000100]

ERSTE ZEILE = 232 ERSTE SPALTE = 231 MIT THRESHOLD = 100,0 UND G-SCHWELLE = 5,0

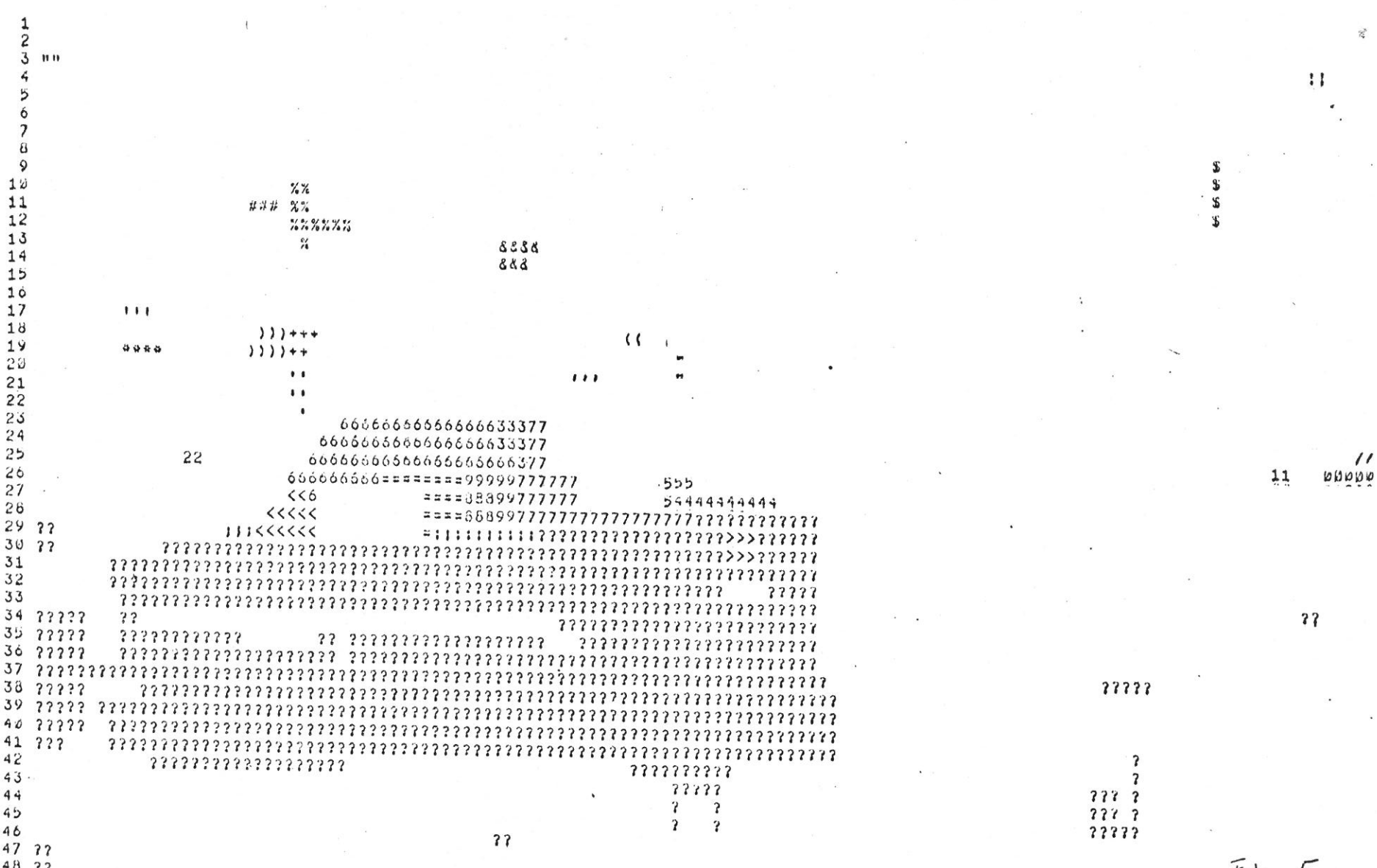

Fig. 5

```
VERSION   V3A41   AM 11-JUN-76 UM 19:20:8,360 UHR  FUER HALBBILD-AUSSCHNITT AUS DSK   I BILD  ,V20 [200330,000103]
ERSTE ZEILE =  232   ERSTE SPALTE =  231   MIT THRESHOLD =  100,0    UND G-SCHWELLE =    2,0
```

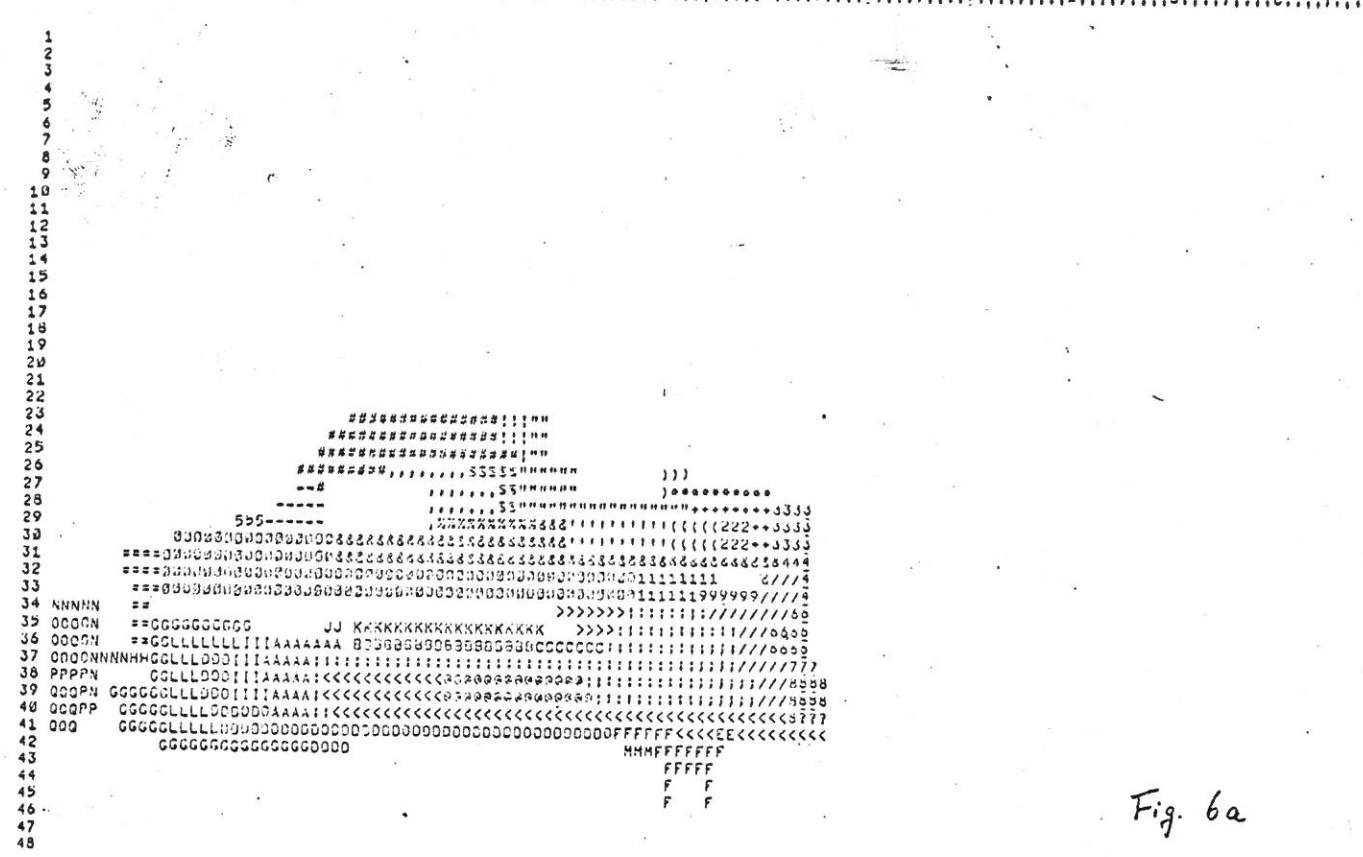

Fig. 6a

```
VERSION   V3A4    AM  8-JUN-76 UM 19:12:22,660 UHR  FUER HALBBILD-AUSSCHNITT AUS DSK   I BILD  ,V07 [200300,000102]
ERSTE ZEILE =  232   ERSTE SPALTE =  231   MIT THRESHOLD =  100,0    UND G-SCHWELLE =    2,0
```

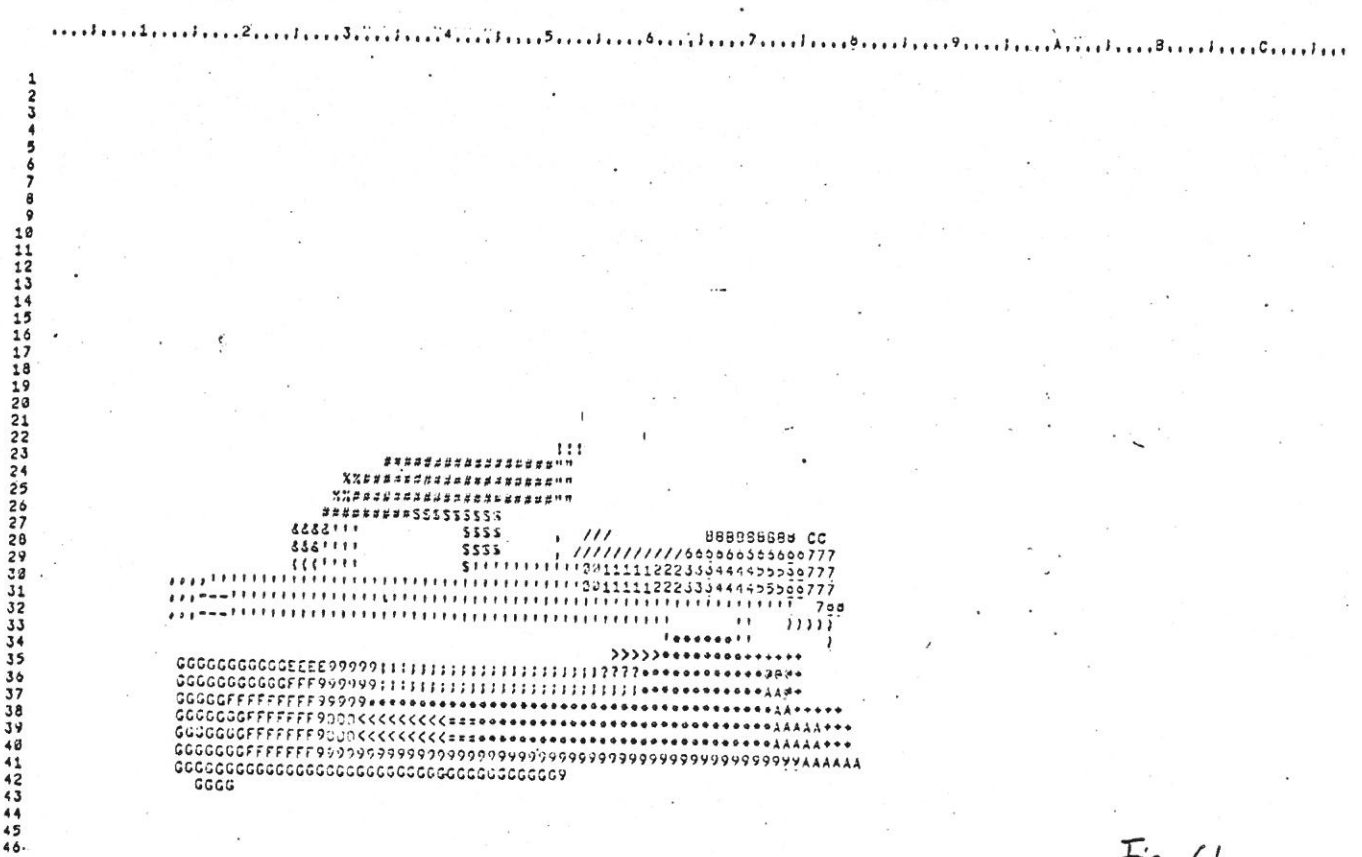

Fig. 6b

VERSION V3A41 AM 11-JUN-76 UM 19:12:1 3,360 UHR FUER HALBBILD-AUSSCHNITT AUS DSK : BILD ,V36 [000336,000100]
ERSTE ZEILE = 232 ERSTE SPALTE = 231 MIT THRESHOLD = 100,0 UND G-SCHWELLE = 2,0

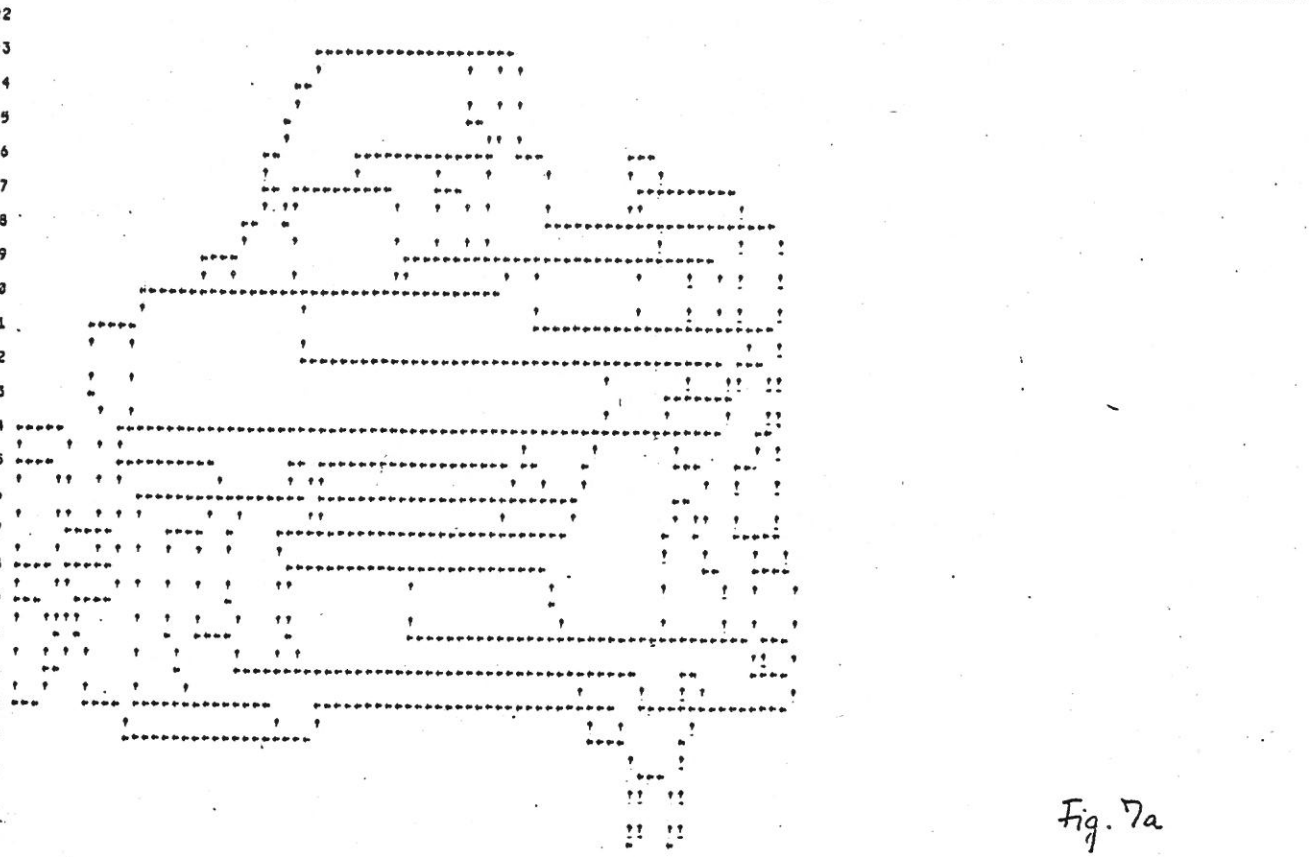

Fig. 7a

VERSION V3A4 AM 8-JUN-76 UM 19:12:22,660 UHR FUER HALBBILD-AUSSCHNITT AUS DSK : BILD ,V07 [000306,000100]
ERSTE ZEILE = 232 ERSTE SPALTE = 231 MIT THRESHOLD = 100,0 UND G-SCHWELLE = 2,0

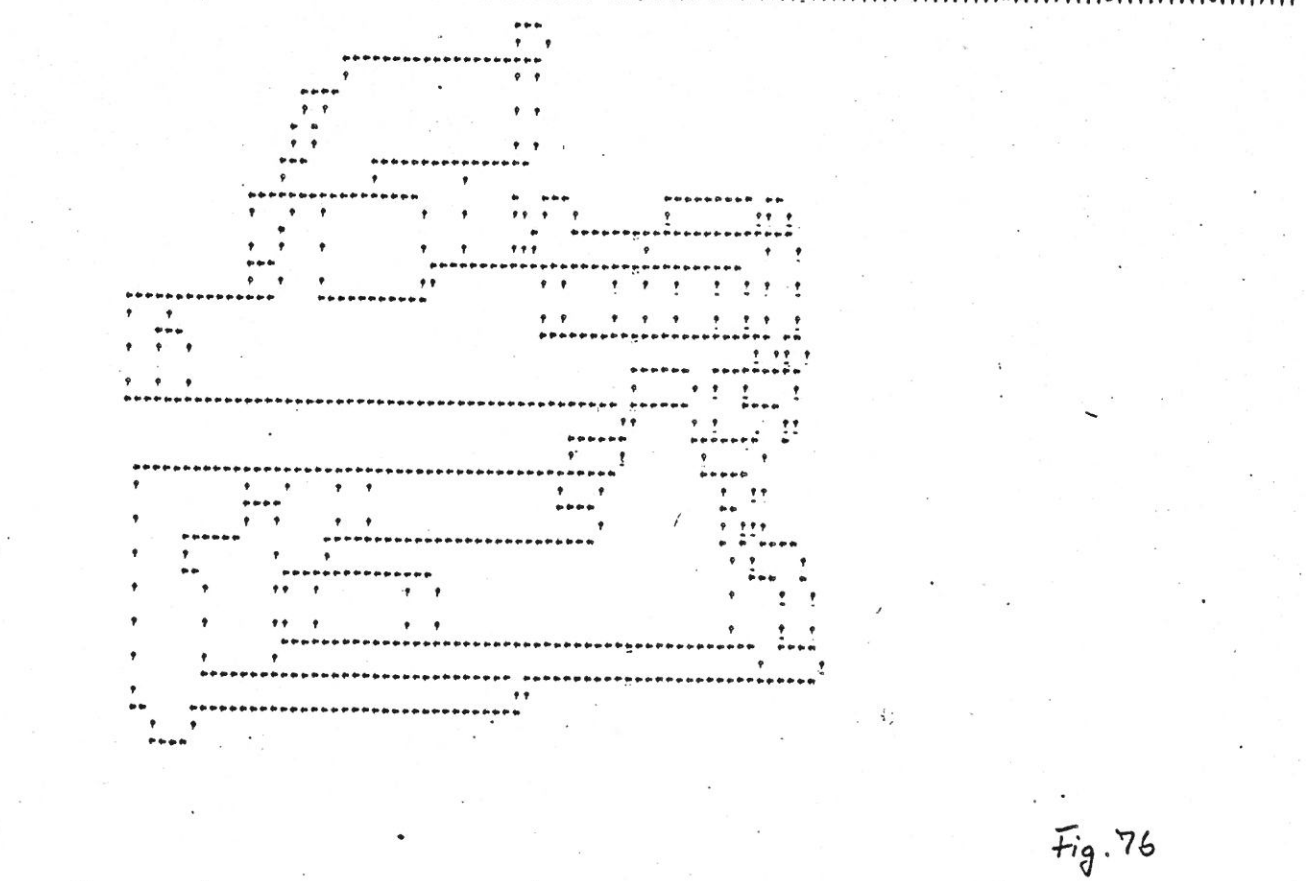

Fig. 7b

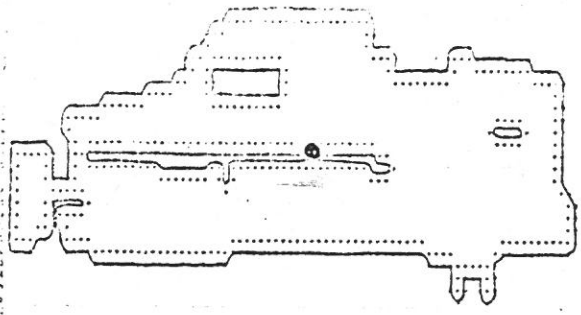

Fig. 9a

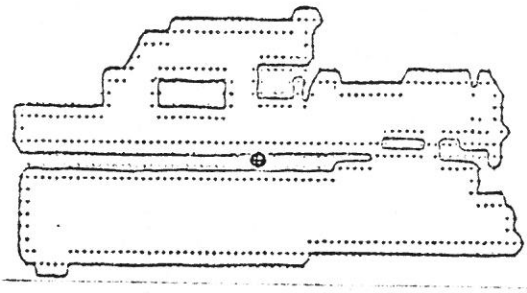

Fig. 9b

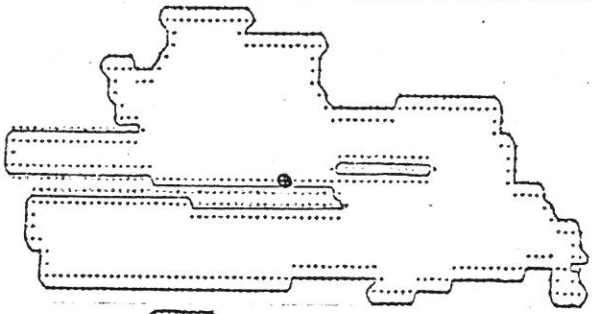

Fig. 9c

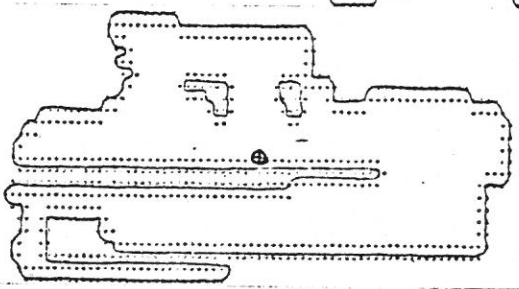

Fig. 9d

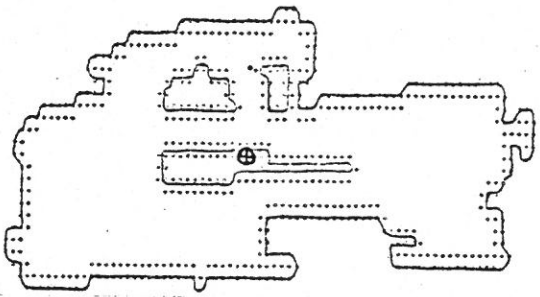

Fig. 9e

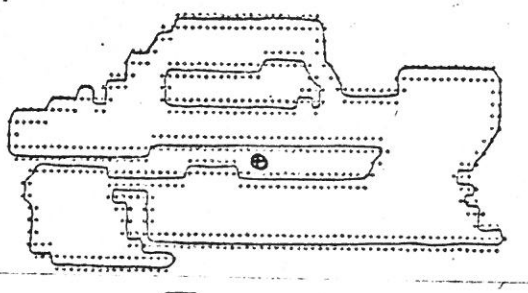

Fig. 9f

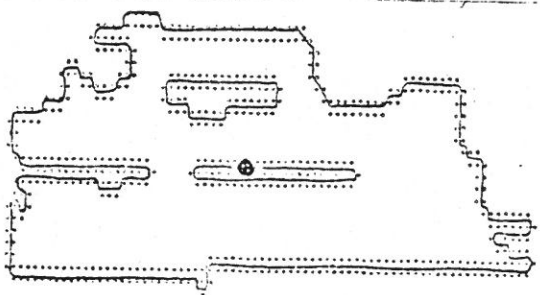

Fig. 9g

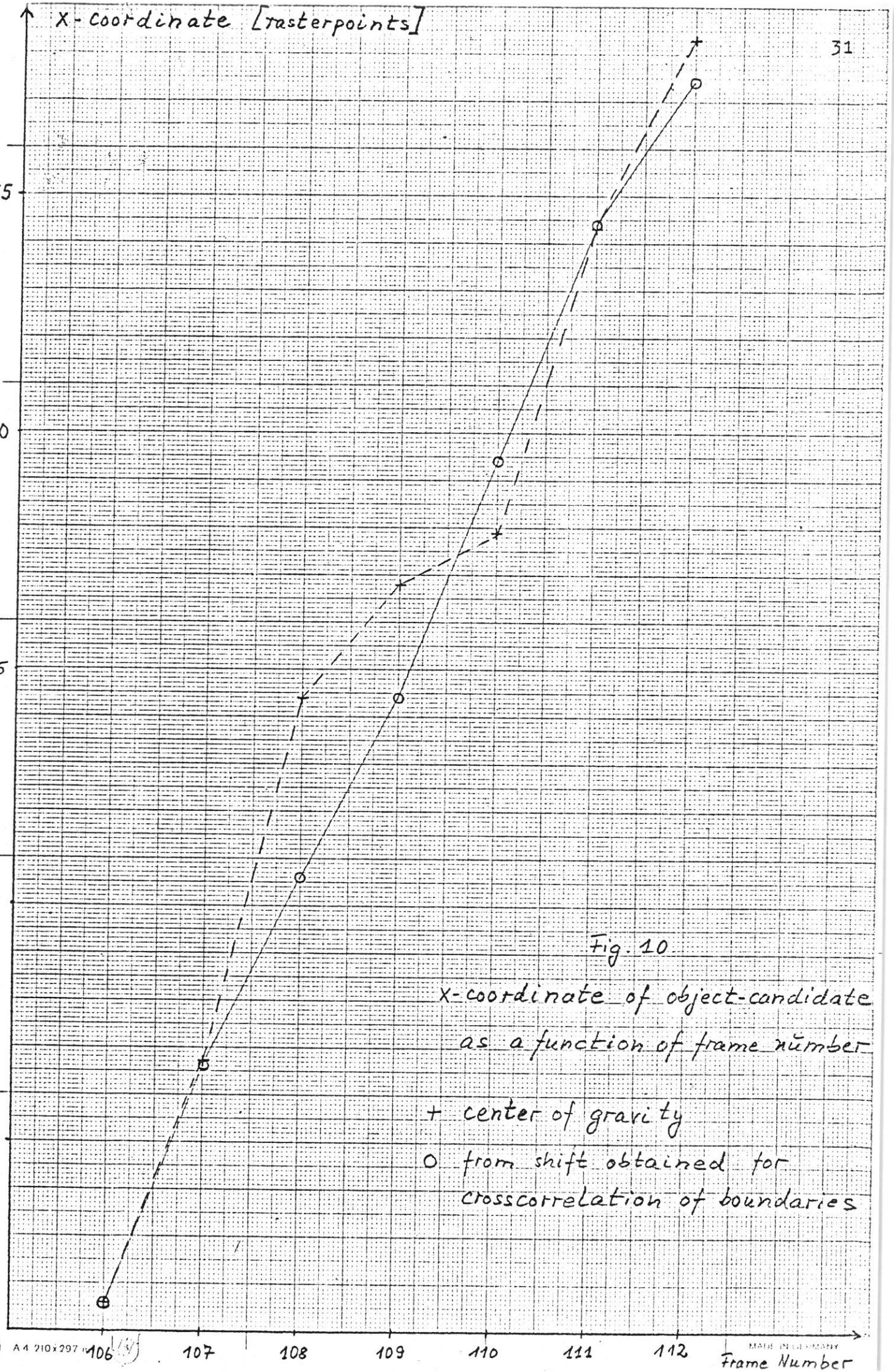

Fig. 10

X-coordinate of object-candidate as a function of frame number

+ center of gravity

○ from shift obtained for crosscorrelation of boundaries

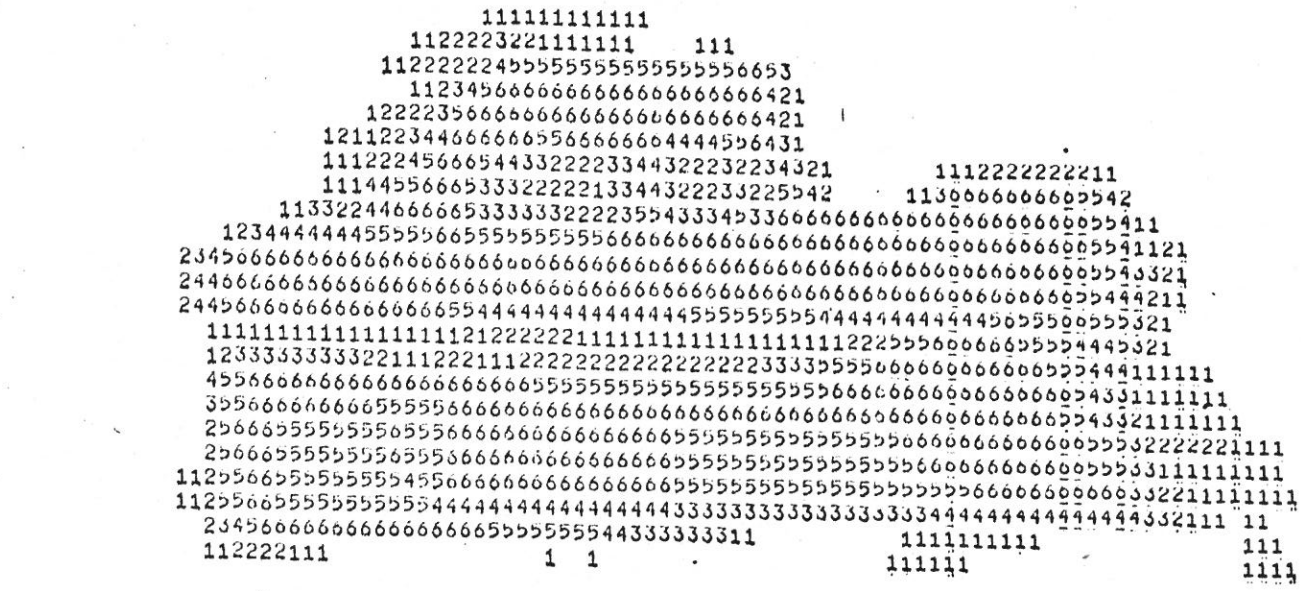

Fig. 11

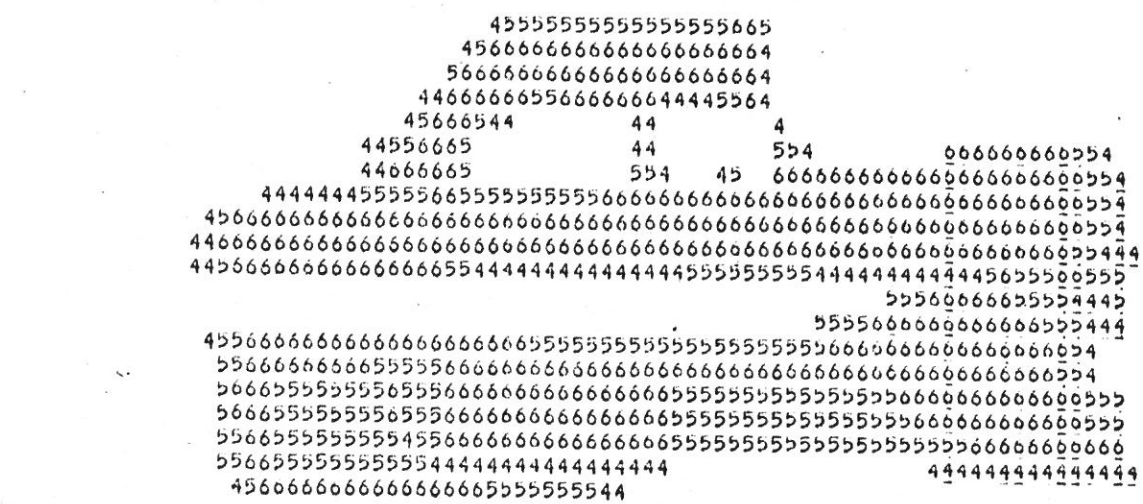

Fig. 12

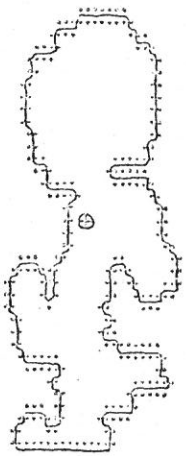

Fig. 16a

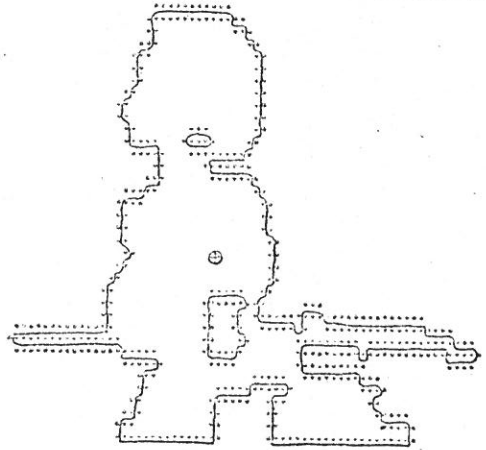

Fig. 16b

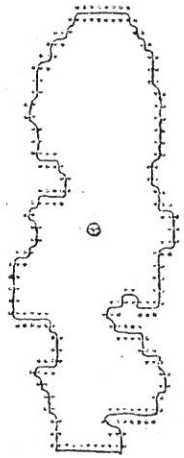

Fig. 16c

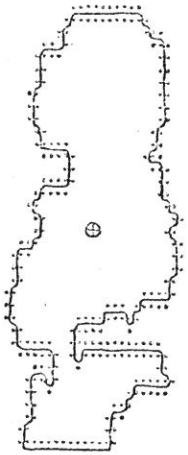

Fig. 16d

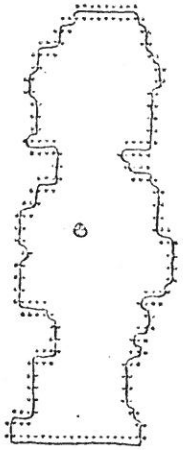

Fig. 16e

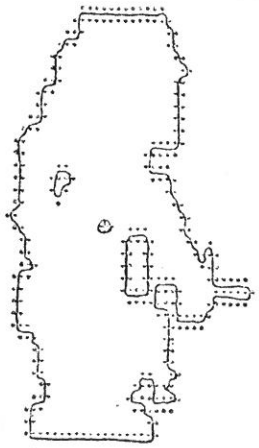

Fig. 16f

Fig. 17

DATUM : 25-JUL-76
UHRZEIT : 17:08:03
BILD1 : BILD V90
BILD2 : BILD V06
SCHWELLE: 10

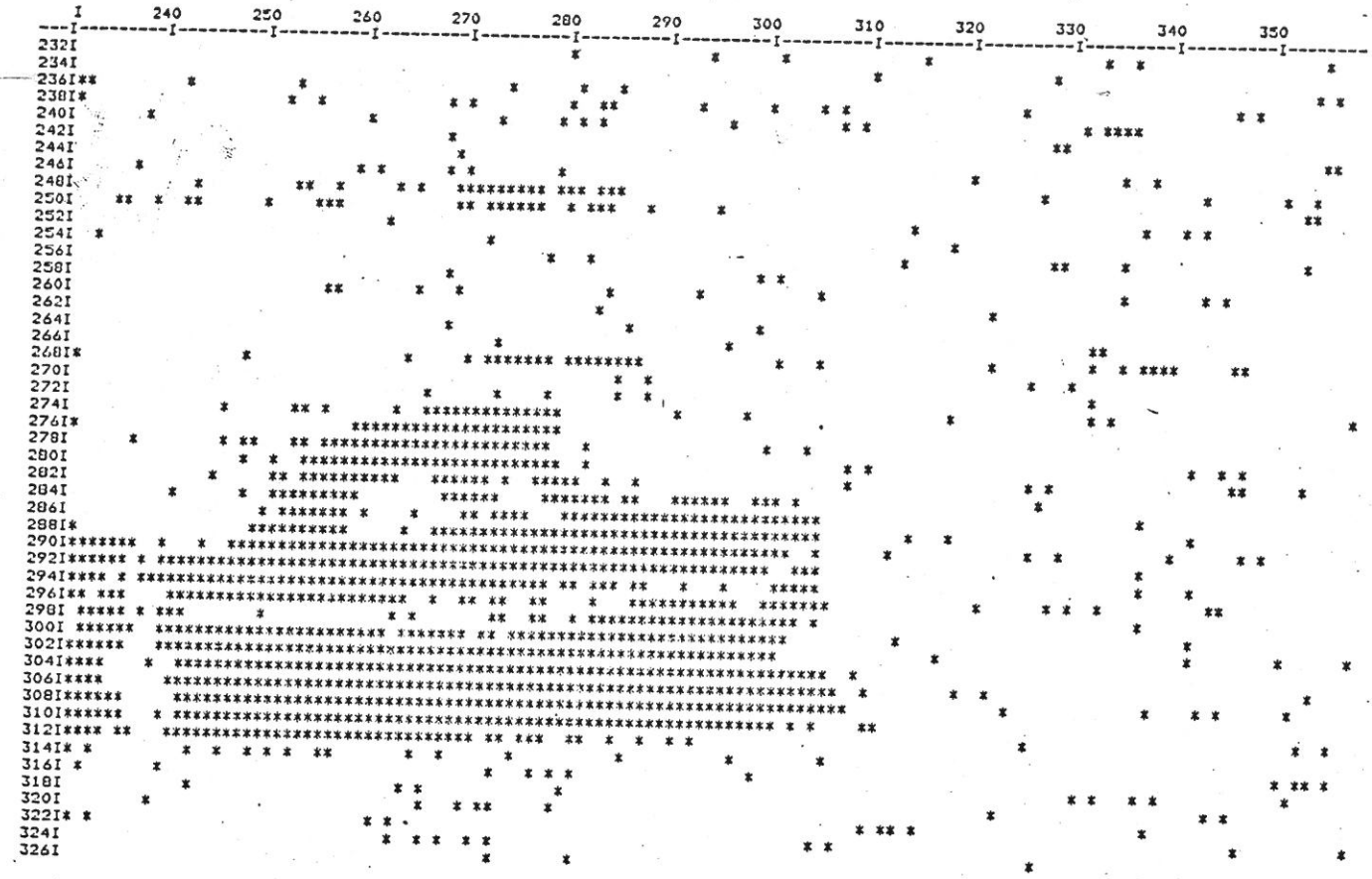

Fig. 19a

DATUM : 25-JUL-76
UHRZEIT : 17:30:43
BILD1 : BILD V90
BILD2 : BILD V06
SCHWELLE: 15

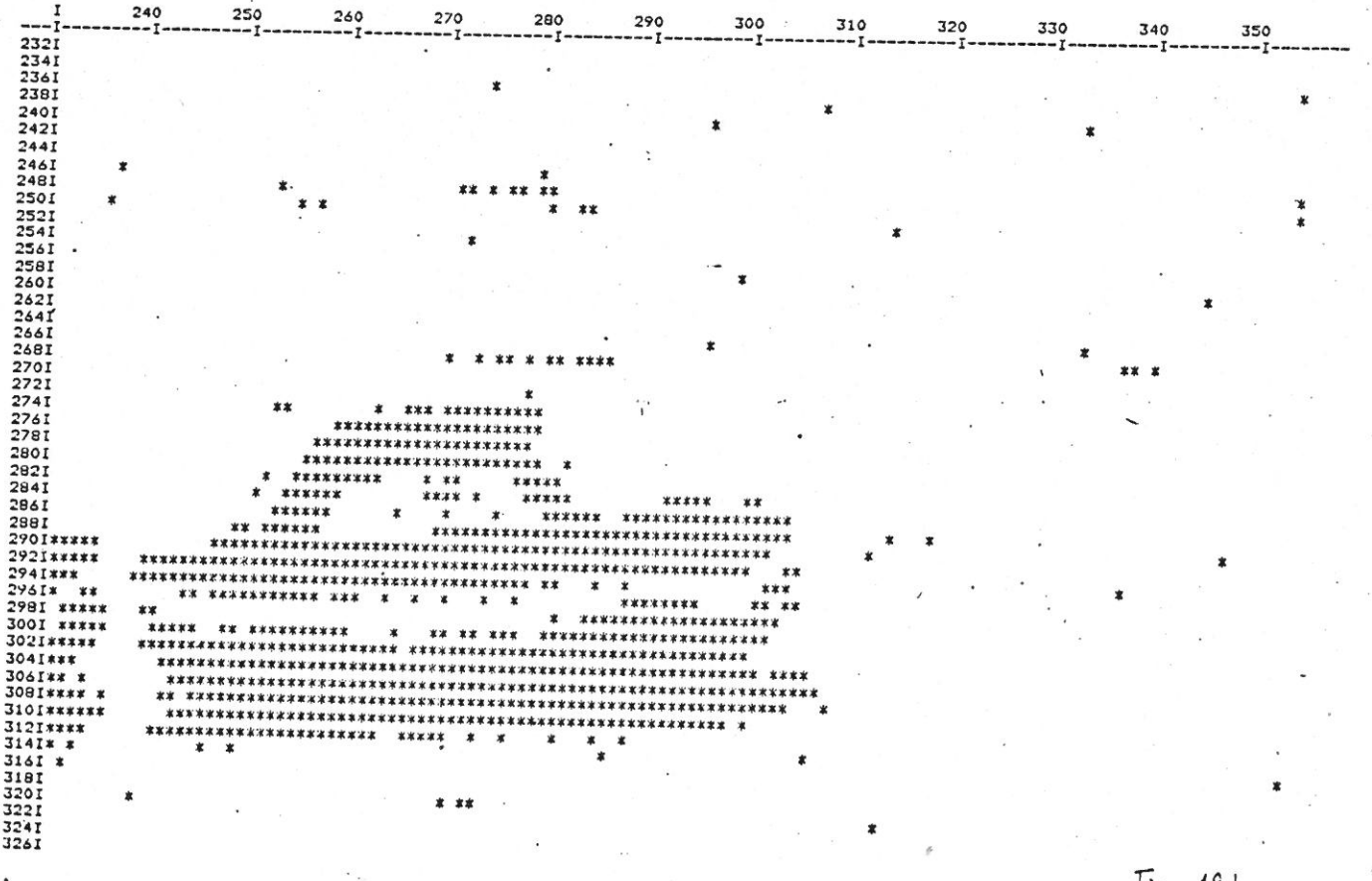

Fig. 19b