

SEPARATING NON-STATIONARY FROM STATIONARY SCENE COMPONENTS IN A SEQUENCE OF REAL WORLD TV-IMAGES

R. Jain (*), D. Militzer, and H.-H. Nagel
Institut fuer Informatik, Universitaet Hamburg
Schlueterstrasse 70, D-2000 Hamburg 13/Germany

Abstract

Results are presented for a new method to identify images of moving objects in a sequence of scene images, e.g. from a TV-camera observing a street intersection. The reported approach exploits the assumption that systematic greyvalue differences - based on second order statistics - between consecutive frames are due to images of moving objects. No knowledge is assumed about size, shape, or texture for images of stationary or non-stationary scene components.

1. Introduction

If a component of a scene is displaced relative to the stationary scene part from moment to moment without significantly changing its internal structure then it appears as a natural abstraction to consider the systematically displaced component to represent a moving object. A sequence of images may, therefore, be evaluated under the assumption that systematic displacements of some image components from frame to frame can be taken as a strong hint to look for the image of a moving object. This assumption opens a way to extract object representations from a sequence of image frames without detailed knowledge about size, shape or textural appearance of the object or the stationary part of the scene. During investigations of this possibility [CNagel 76b, Nagel 77] it became necessary to quickly and reliably identify non-stationary image components in a sequence of images, e.g. TV-frames from a TV-camera observing a street intersection with cars and pedestrians or a conveyor belt with industrial parts.

We want to suggest an approach which seems to offer a reliable basis for separating non-stationary from stationary image components.

2. Comparison of digitized images based on second order statistics

A complete digitized TV-frame comprises of 573 lines with 512 pixels each at our installation. The area consisting of four neighbouring pixels for six consecutive lines is defined as a 'geo-pixel'. For each geo-pixel the mean greyvalue m and the corresponding variance s is calculated and stored. Since the greyvalues are digitized to eight bits and - due to computational reasons - not the mean

and variance but the sum of greyvalues and squared greyvalues are preserved for each geo-pixel, the amount of storage required for each geo-pixel is given by 5+8 bits for the greylevel sum and 5+16 bits for the sum of the squared greylevels. Since these sums will have to be accumulated for several frames, even more bits have to be reserved. By suitable packing, these data fit - for each geo-pixel - into two 36 bit words of our DECSys-10. We thus require $2 \times 96 \times 128 = 24576$ words of core for an entire TV-frame in geo-pixel format.

For every TV-frame in the sequence, each geo-pixel is compared to data previously observed at the same geo-pixel coordinates by means of the following

$$\left[\frac{\bar{S} + S_{n+1}}{2} + \left(\frac{\bar{m} - m_{n+1}}{2} \right)^2 \right] \frac{2}{S * S_{n+1}}$$

m and s denote the mean greyvalue and its variance for the measurements accumulated at this geo-pixel position. This expression is formed in analogy to one given by Yakimovsky 76 to determine whether two neighbouring test areas can be thought of being measurements from identical or from differing normal distributions for greylevels. Here, the measurements are not taken from two neighbouring areas of the same TV-frame but rather from the same area of two neighbouring TV-frames - an idea which comes quite natural when working with the Yakimovsky algorithm [CYakimovsky 76, Nagel 76a3] on sequences of TV-frames [CNagel 76b3].

If the likelihood ratio turns out to be smaller than a threshold t (in our experiments chosen within the range 1 to 20) then the two sets of measurements for this geo-pixel position are considered to be drawn from the same normal greylevel distribution. On this basis, the sum of greylevels as well as the sum of squared greylevels for this geo-pixel from frame $n+1$ are added to the corresponding, already accumulated sums for the geo-pixel at the same coordinates and the count of contributing pixels for this geo-pixel is increased by 24. This corresponds to taking repeated measurements and averaging them to obtain a better estimate for the grey value of this geo-pixel.

If, however, the likelihood ratio equals or exceeds the chosen threshold t , it is decided to attribute the measurements for this geo-pixel in frame $n+1$ to a different normal greylevel distribution than the one from which the hitherto accumulated measurements for the same geo-pixel coordinates were obtained. In this case the sums of greylevels and squared greylevels are entered together with the frame number and the geo-pixel location into a list

(*) On leave from Electronics & Electrical Communication Engineering Department, Indian Institute of Technology, Kharagpur 721302, India

of (yet unconnected, see later) 'not matched' geo-pixels which is ordered according

- first to line numbers,
- for each line according to column numbers,
- for each column number according to frame numbers.

Figure 1 shows two pictures of a TV-frame sequence from a street intersection scene. Figure 2 shows the numbers of geo-pixels at each of the 96x128 geo-pixel positions which are in the 'not matched' or 'failing match' list when seven subsequent frames are compared in the described manner. One can easily see areas with a high density of 'not matched' geo-pixels. With the exception of the vertical band at the left which is due to hardware trouble (line jitter compounded by some other effect), one can attribute the high density 'not matched' areas to moving objects: the bright car crossing the intersection, a pedestrian to the lower left and a car at the lower center coming to a stop.

If the failing match of a geo-pixel is due to genuine motion of the corresponding object then the next frame should also not match at the same geo-pixel position. This consideration is used to build a filtering process by which the accidentally failing matches can be removed to some extent with a local operation.

As soon as a failing match is observed at a geo-pixel position in frame $n+1$, this position is flagged. If comparison of the geo-pixel from frame $n+2$ with the data accumulated at the same coordinates in the geo-pixel matrix up to frame n yields a failing match, too, then both failing matches are definitely accepted. This situation is henceforth denoted as a mismatch. Moreover, this geo-pixel position continues to be flagged until a match will be observed for it. If the geo-pixel from frame $n+2$ matches to the data accumulated up to and including frame n then the failing match at this position in frame $n+1$ is dropped. The corresponding greyvalue and squared greyvalue data from frame $n+1$ are not incorporated into the accumulated measurements, whereas those of frame $n+2$ are incorporated. The 'failing match' flag is removed for this geo-pixel position. Figure 3 presents the definite mismatches obtained by comparing the same TV-frames that resulted in figure 2. Most of the noise has disappeared.

Based on these results, two alternative approaches have been devised to extract non-stationary image components.

3. Clustering of mismatched geo-pixels

If a mismatch for a geo-pixel position could really be attributed to definite changes in greylevels for this position due to motion of an object in the scene then sooner or later mismatches must be observed in neighbouring geo-pixels. Therefore, a candidate for a non-stationary image component is formed for every 4-connected group of mismatched geo-pixels which

- contains at least one geo-pixel for which the number of mismatches exceeds a certain fraction of the number of TV-frames compared (this fraction is currently taken to vary between 50 % and 85 %): a so-called 'strong mismatch',
- contains at least one additional mismatched geo-pixel which is 4-connected to a 'strong mismatch'.

Additional mismatches may be 4-connected to a 'strong mismatch' either directly or indirectly through other mismatches. It is considered sufficient to look for all 4-connected mismatches since even for oblique motion relative to the TV-raster enough neighbouring geo-pixels should eventually show a mismatch to establish 4-connection between them.

Figure 4 presents the candidates for non-stationary image components extracted from the data represented in figure 3.

Up to this point the algorithm contains only three parameters: the threshold t for the likelihood ratio, the number N of frames after which the list of mismatched geo-pixels is searched for non-stationary component candidates, and the fraction of N which is required for a strong mismatch. One could consider the minimum number of 4-connected mismatches in a cluster as another parameter which for some pictures might be set higher than 2 as it is done here. Now an additional step is introduced which provides the setup to repeat the steps discussed so far.

An enclosing rectangle for each non-stationary image component candidate is determined. One pair of rectangle sides is taken to be parallel to the scanline. The coordinates of the rectangle edges are remembered for a later estimate of the velocity vector to be attributed to the non-stationary component candidate. After these preparations the next N frames are compared to the geo-pixel matrix. Whenever a mismatch occurs at a geo-pixel position 4-connected to a candidate as enclosed within the (latest) rectangle, then this mismatch is considered to belong to the candidate to which it is u-connected (note that a mismatch requires at least two subsequent 'non-matches' and that it must be 4-connected to a candidate for the non-stationary image component itself, not to the rectangle enclosing it). If a mismatch cannot be 4-connected to an already existing non-stationary component candidate, it is introduced in a separate list of 'yet unconnected' mismatches. In this way even mismatches due to very slow motions will eventually be recognized after sufficient frames have been compared.

Whenever another series of N frames has been treated in this manner, the enclosing rectangles about the newly added mismatches of already existing non-stationary component candidates are determined. Then the left and right edges of such a newly enclosing rectangle are compared with those of the preceding enclosing rectangle of the same candidate. The larger of the two edge displacements for each candidate will determine an estimate of the image displacement velocity for this candidate

along the horizontal axis. The smaller displacement velocity is attributed to residual mismatches in the trailing part of the non-stationary image component candidate. To take this hypothesis into account, the trailing fraction of the newly determined enclosing rectangle equal in extension to the displacement difference between leading and trailing edge of the enclosing rectangle during the last N frames is released. (Note that this implies assuming a constant extension of the non-stationary image component candidate along the apparent direction of motion.) As a consequence, all entries for these geo-pixel positions are cleared since it has just been decided that the non-stationary image component moved away from there and uncovered the background. New estimates of the background can be accumulated in these geo-pixel positions during subsequent frames. The same approach is taken with respect to the vertical displacement. Before handling the next N frames, the list of 'yet unconnected' mismatches is searched for additional non-stationary image component candidates; all remaining isolated mismatches are then thrown away.

This algorithm has been tested on four image sequences from different street scenes, each containing 25-50 TV-frames. The threshold t had to be varied between 3 and 20 to yield optimal results for these image sequences. The number N of frames in one subseries has always been chosen to be 7 and the fraction of it establishing a strong mismatch was best taken to be 85 % (6 out of 7 frames).

In order to obtain good results for all four image sequences with the same threshold value, the match criterion of paragraph 2 has been modified by replacing all variances s with $\text{MAX}(s, 20)$ - note that the nominator of the match criterion represents the square of the variance for greyvalues from all pixels involved in this test.

About 8-9 seconds CPU-time of a DEC KI-10 processor are required to test already prepared geo-pixel data of one TV-frame for a match with previously accumulated data. If one starts directly from the raw digitizings, about 25-30 seconds per frame are required. After each subseries of N frames an additional 10 seconds CPU-time are required to process the mismatches found in this subseries (about 500-3000 per frame depending upon the threshold value).

The final extraction of moving object descriptions from non-stationary image components may proceed according to the methods discussed in Nagel 76b and Nagel 77.

4. Exploiting a monotony characteristic of the mismatch count

Based on the mismatches established by the algorithm of paragraph 2 an alternative approach to the one described in paragraph 3 has been devised. This alternative may either be used to complement the one described in the preceding paragraph or to yield another estimate for the image of a moving object, thus allowing a consistency check on the

estimation of the image of the moving object.

The differencing operation described in section 2 results in clear indication of those regions of the frames where changes are taking place. This operation offers one very attractive property which may be exploited for extracting the image component potentially due to a moving object. Let us consider an idealized situation to understand this property: an object with homogeneous greyvalue is moving parallel to the image plane of the TV-camera against a homogenous background. At the rear end of the object some part of the background which was covered by the object in the previous frame is uncovered and at the front end some uncovered part of the background is covered by the moving object. This results in mismatched geo-pixels. Assuming the reference frame to be represented as 0th frame, after the 1st frame there will be 1 entry for each 'not matched' geo-pixel corresponding to these regions. The 'not matched' geo-pixels may be thought of as representing a difference picture.

After the second frame, the regions of the difference picture which were having 1s become 2s and the new regions due to the uncovering and covering of background will have 1s. After the 3rd frame, 2s become 3s, 1s become 2s and newly created regions receive 1s. If the object motion is unidirectional, this process continues until the object image has moved over the distance equivalent to its projection along the direction of motion in the image plane of the TV-camera. This phenomenon is depicted in figure 8 for the comparison of 4 frames. Hence if one observes neighbouring entries in the difference picture corresponding to this idealized situation, the entries are found to be monotonic in nature. The above discussion is valid for the motion having a velocity component perpendicular to the line of sight of the camera. If the velocity component in the image plane of the TV-camera is zero then mismatches for a motion along the line of sight of the camera will be due to size changes for the projected image of the moving object. However, this case is not considered in the present implementation.

In real world pictures the situation is not, unfortunately, so straightforward. Because of the inhomogeneous objects and inhomogeneous background, some noise is introduced in the difference picture. Still it should be expected that for the most part these entries (assuming the noise not to be very high, in high noise situation even the human beings make mistakes) will confirm monotonicity. Accordingly, for an object having a velocity component perpendicular to the axis of the camera there should be two more or less monotonic regions - provided this object is completely contained in the field of view, is not occluded, and does not show vanishing contrast with respect to the stationary scene part that is currently occluded by this object.

The monotonicity of the entries will also indicate the direction of the motion: the object will be moving in the direction in which the entries are decreasing. On knowing the direction of the motion

it is immediately known which region corresponds to the rear end of the object and which to the front of the object. It should be noted that the mismatch region at the rear end of the object is formed due to uncovering of the background while the mismatch region at the front end is formed due to covering of the background by the object.

Hence the mismatch regions in the n th frame correspond to the background component and the object component, respectively. These two regions may be utilized to estimate a representation of the non-stationary image component.

In the present algorithm we extract the properties of the background (presently only greylevel) and then build up the non-stationary component by considering that all the pixels in the area fixed by the two regions of the difference picture which are differing from the background by an appreciable amount represent the object. This algorithm is applied to the TV-sequence of the traffic scene. The result obtained from frame 106 is shown in figure 6. A simple filtering of noise is achieved by removing all the 4-connected regions of size less than 24. The regions of the non-stationary component having different greylevels are printed using different characters in the filtered non-stationary components shown in figure 7. It should be noted that in this example the car and its shadow are considered to be the same non-stationary image component because they are moving together, although they have very different grey-values.

Once these non-stationary image components are extracted, estimation of their velocity components presents no problem. Moreover, the representation of the non-stationary image component candidates may be compared with the representation of the corresponding candidates in subsequent frames. Any pixels which do not appear in a large fraction of candidates will be discarded from the representation of the non-stationary image component. Additional domain independent knowledge may subsequently be applied to support the hypothesis that the non-stationary image component can indeed be taken as a representation for the projected image of a moving object in the scene. However, such techniques exceed the scope of this contribution.

5. Conclusion

The experiences with Yakimovsky's algorithm show that a segmentation of greylevel pictures based on the above mentioned second order statistics likelihood ratio seems to be a very robust method Nagel 76a. Further experiments are necessary to verify whether application of the same approach to the comparison of consecutive frames turns out to be robust, too.

The method proposed by Potter 75 seems to present difficulties if applied to real world scenes. Chow and Aggarwal 77 work with objects showing a high contrast against background so that noise problems do not bother them. Part of the results reported by Chien and Jones 75 refer equally to experiments

with well controlled contrast. Their work on real-time tracking of cars in a street scene is primarily concerned with finding and following prominent greyvalue features rather than extracting a description of a moving object from the TV-frame sequence as in our work. Limb and Murphy 75 developed a method for the detection of a non-stationary image component in the context of bandwidth compression of TV-signals. They essentially rely upon a threshold for the greyvalue difference between successive frames to find a non-stationary image component. They do not attempt to isolate and extract the image of a moving object. Further references to the literature may be found in Nagel 76b and Nagel 77.

It is obvious that the combinations of the algorithms proposed here with evaluation of segmentation results for individual images may yield additional information as to what regions are likely to belong to the image of a moving object. Especially in the case where the leading and trailing edges characterized by mismatches can be connected within each frame by (a group of) regions with constant greyvalue characteristics, strong support would be obtained to consider these (group of) regions as a representation for a moving object.

It might be worthwhile to point out that no special emphasis is placed by this approach on the first frame. Although it provides the starting estimates for each geo-pixel characteristic, these are either reenforced by accumulation of data from compatible geo-pixels in later frames - indicating that this geo-pixel belongs to a stationary image component - or new estimates are started once the non-stationary component has disappeared from this area of the image.

It should be noted that the approach suggested here will allow to determine automatically a window within each image around a suspected moving object candidate. In addition, it can be used to derive the frame number difference between two frames within which the non-stationary image component has been displaced by more than its own extension along the direction of motion. Therefore, a direct comparison of these two frames will contrast the non-stationary image component against parts of the stationary image component in the other frame, thus enabling a simple extraction. The automatic determination of a (conservative) window around the non-stationary image component and the appropriate frame number difference are the essential parameters for the method described in Nagel 76b where these parameters still had to be estimated by a human operator and supplied as input values. Therefore, the current contribution can be seen as giving additional support to the approach described in Nagel 76b. It should be emphasized that the approach described here must be seen as one part in an entire system for analysis of image sequences, as e.g. in the envisaged design of a system for the analysis of TV-frame sequences outlined in Nagel 77.

6. References

- Chien and Jones 1975, Acquisition of Moving Objects and Hand-Eye Coordination, p. 737, Advance Papers IJCAI 4, Tbilisi, Georgia/USSR, Sept. 3-8, 1975
- Chow and Aggarwal 1977, Computer Analysis of Planar Curvilinear Moving Objects, IEEE Trans, on Comp. C-26, 179 (1977)
- Limb and Murphy 1975, Estimating the Velocity of Moving Images in Television Signals, Computer Graphics and Image Processing vol. 4, 311 (1975)
- Nagel 1976a, Experiences with Yakimovsky's Algorithm for Boundary and Object Detection in Real World Images, Proc. IJCP3, Coronado/Cal., Nov. 8-11, 1976, p. 753-758
- Nagel 1976b, Formation of an Object Concept by Analysis of Systematic Time Variations in the Optically Perceptible Environment, IfI-HH-B-27/76 (July 1976), Institut fuer Informatik, Universitaet Hamburg; to appear in Computer Graphics and Image Processing
- Nagel 1977, Analysing Sequences of TV-Frames: System Design Considerations, IfI-HH-B-33/77 (March 1977), Institut fuer Informatik, Universitaet Hamburg; condensed version appears at IJCAI-77
- Potter 1975, Scene Segmentation by Velocity Measurements Obtained with a Cross-Shaped Template, p. 803, IJCAI 4, Tbilisi, Georgia/USSR, Sept. 3-8, 1975
- Yakimovsky 1976, Boundary and Object Detection in Real World Images, J. ACM vol. 23, 599-618 (1976)

7. 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 TV-camera.

Fig. 2: 'Not matched' or 'failing match' geo-pixel positions from comparison of frame 90 through 96. The digit at each geo-pixel position indicates how often a 'failing match' occurred for that geo-pixel.

Fig. 3: The digit at each geo-pixel position indicates the number of definite mismatches after applying the local filter criteria described in paragraph 2.

Fig. 4: Candidates for non-stationary image components extracted from the data shown in fig. 3 according to the procedure described in paragraph 3. The different digits correspond to different candidates. All geo-pixel positions attributed to a candidate for a non-stationary image component are indicated by the corresponding digit.

Fig. 5: aa', bb', cc' and dd' represent the suc-

cessive positions of a moving object image in the 0th, 1st, 2nd, 3rd frame, respectively. The numbers above the leading and trailing part of the moving object image indicate the number of failing matches determined after comparing the 1st, 2nd, 3rd frame with the 0th frame.

Fig. 6: The object components extracted from frame 10b using the approach described in paragraph 4,

Fig. 7: The regions of the components shown in fig. 6 are reproduced in fig. 7 with the regions of different greyvalues being represented by different characters.

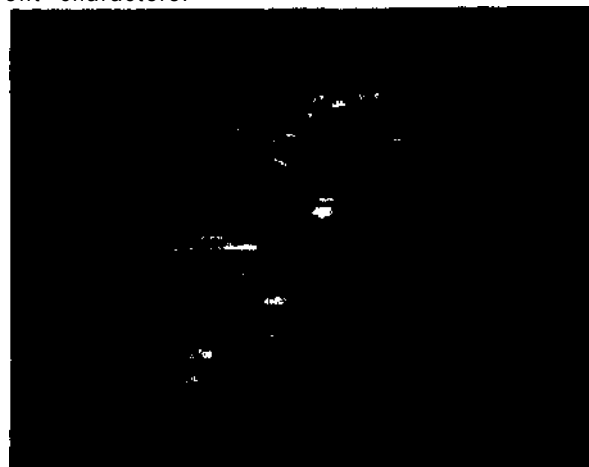


figure 1a

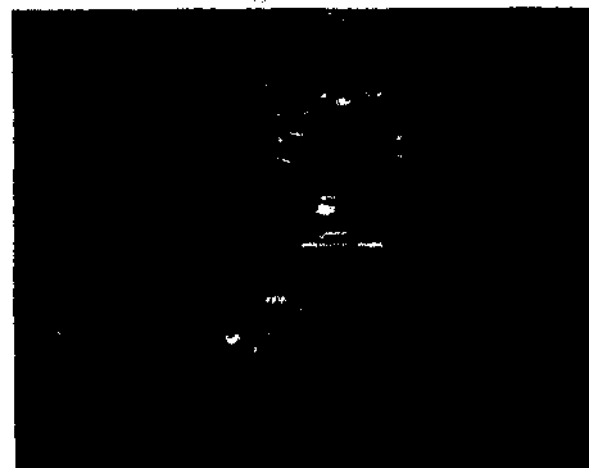


figure 1b

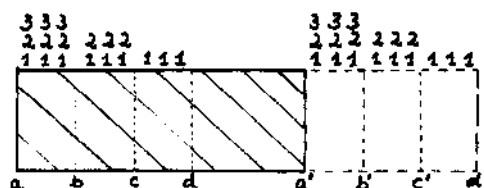


figure 5

Fig. 6

Fig. 7

Fig. 4

Vision-4: Jain
618