Pattern Recognition Vol. 18, Nos 3/4, pp. 207–213, 1985.
Printed in Great Britain.

0031-3203/85 $3.00+.00
Pergamon Press Ltd.

DETECTION OF THE MOVEMENTS OF PERSONS FROM A SPARSE SEQUENCE OF TV IMAGES

TOSHIFUMI TSUKIYAMA and YOSHIAKI SHIRAI
Electrotechnical Laboratory, 1-1-4, Umezono, Sakura-Mura, Niihari-Gun, Ibaraki, Japan

(*Received* 17 *February* 1984; *received for publication* 2 *May* 1984)

Abstract—This paper describes a method for detecting persons walking in a passageway from two consecutive TV images at long intervals. The method consists of the following steps: (1) extracting candidate areas for persons from TV images by image processing; (2) detecting persons in the areas on the basis of their shapes and sizes, and locating them in a passageway; (3) finding the correspondence of persons between two images on the basis of their locations. Special hardware called DIP was employed for high speed image processing.

Motion estimation Robot vision Movement of persons Autonomous vehicle
High speed image processing Consecutive TV image Inverse perspective transformation
Occlusion problem

1. INTRODUCTION

Motion estimation of 3-D objects from a sequence of TV images has an application in many areas, included target tracking and robot vision. Motion estimation of 3-D objects consists of two steps: extracting image-space shifts of objects from a sequence of images and determining the motion parameters from these shifts. In the second step the equations obtained from a camera model are solved, based on these shifts. A number of papers[1] have been presented for detecting the image-space shifts of objects from images. There have been three basic approaches to the detection problem: the matching method; the temporal-spatial gradient method; the difference method.

The matching method is a straightforward approach in obtaining image-space shift of objects. A set of prominent features in one image is found and subsequent images are searched for the corresponding features. Prominent features are, for example, the image points which might express corners of objects.[2] The image points are matched between two images, based on the assumption that the movements of objects are very small between two consecutive images. Then, assuming that objects are rigid, the groups of image points on the same object are found.

The temporal-spatial gradient method[3, 4] uses the velocity components of image points, which can be calculated from the spatial gradient of the images and the local intensity change over time due to motion. The frame rate is assumed to be sufficiently rapid such that the value of the gradient at an image point does not change significantly between two consecutive images. The image points on the object are estimated using constraints of the movements, such that neighbouring points on the object have similar velocities and the apparent velocity of the brightness patterns in the image varies smoothly almost everywhere.[4]

The difference method[5] compares a current frame with a reference frame to find areas of large difference in brightness. Then each area is tracked through a sequence of images and merged to determine the surface boundary of a moving object. The movement of an object from frame to frame is assumed to be small, so that an area corresponds to a portion of a moving object.

Since those approaches mentioned above require similarity between two consecutive images, it is assumed that objects are rigid or that the interval between two images is very short, say TV frame rate. From the viewpoint of applications to robot vision, real time image processing is desirable. Since it requires much computation to process the image data of complex scenes, the interval of two input images cannot be very short. In addition, flexible objects must be dealt with in order to detect movements of persons. In this paper we propose a method for detecting movements of persons walking in a passageway from a sparse sequence of TV images. The detection of movement is to find the trajectories of persons on a flat plane. In our method (shown in Fig. 1) the objects are at first extracted from each image and the locations of the objects are calculated in the passageway. Correspondences of objects between two images are found by comparing the location changes. While the conventional matching method finds the correspondence of image points between two images, our method finds the correspondence of objects themselves. Thus the latter method is effective even if the size and shape of a moving object changes to a great extent. The method

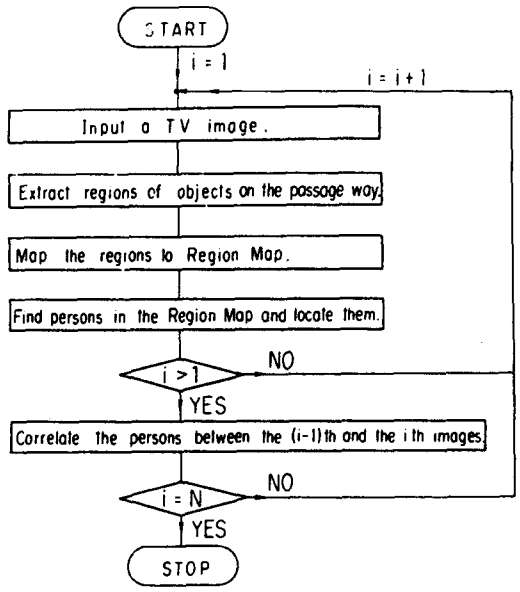

Fig. 1. Flow chart of the method.

will be applied to a vision system for an autonomous vehicle with a TV camera, such as one for guiding the blind in a building.

Here we deal with a simplified case where the direction of a TV camera is fixed and only persons are in a passageway. Image data of a passageway scene are taken from a TV camera, as shown in Fig. 2. The process proceeds as follows: (1) finding candidate areas for persons in images, based on the mean and the variance of brightness; (2) detecting persons in the areas on the basis of their shape and size and locating the persons in the passageway; (3) finding the correspondence of the persons between two consecutive images on the basis of their locations. These steps are described in detail in Sections 2, 3 and 4, respectively. In order to catch up with the movements of persons, image data should be processed very fast. We used a special hardware for the high speed processing. The details are discussed in Section 6.

2. SEGMENTATION OF IMAGES

This section describes a method for finding candidate areas for persons in each image. Since the width of the passageway and the direction of the TV camera

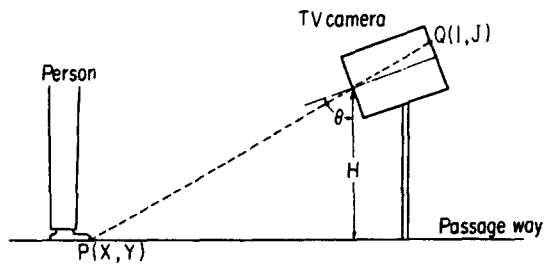

Fig. 2. Range finder with one TV camera.

are known, side parts of a passageway, such as wall areas, are extracted from the image in advance. The remaining areas are split into the following two regions; (1) a region corresponding to the floor of the passageway; (2) a region corresponding to objects (persons) on the floor.

The passageway is assumed to be illuminated by many lamps on the ceiling and the reflectance of the floor of the passageway is large enough. Under these conditions, vertical surfaces in the passageway are observed as being much darker than the horizontal floor up to a certain height. Consequently, the mean of brightness of persons becomes smaller than that of the floor of the passageway. However, some parts of person's clothes may be a little brighter than the floor, due to wrinkles. In this case, the variance of brightness of the parts is expected to be larger than that of the floor. We use the mean and the variance of brightness of image points to find the floor region and the object region.

Let $I(i, j)$ denote the brightness of the pixel at (i, j) in an image. We use the following measures for the mean $M(i, j)$ and the variance $D(i, j)$ of the brightness of the pixel at (i, j)

$$M(i, j) = \sum_{m=-1}^{1} \sum_{n=-1}^{1} I(i + m, j + n)$$

$$D(i, j) = \left| \sum_{m=-1}^{1} [I(i + m, j + 1) - (i + m, j - 1)] \right| + \left| \sum_{n=-1}^{1} [I(i + 1, j + n) - I(i - 1, j + n)] \right|.$$

Each pixel is classified into the floor region or the object region by the following procedures.

(1) If $M(i, j) > I1$, then it is classified into the floor region.

(2) If $M(i, j) < I2$, then it is classified into the object region.

(3) If $I1 > M(i, j) > I2$ and $D(i, j) < D1$ then it is classified into the floor region.

(4) Otherwise, the pixel is classified into the object region.

$I1$, $I2$ and $D1$ are constant values and fixed based on the first TV image.

The object region is usually divided into multiple connected regions, each of which is denoted by 'object area'. Finally, if there are small object areas surrounded with floor area, they are merged with the floor region. Figure 3 shows examples of an input image and the floor region of the image.

3. FINDING PERSONS

This section describes a method for finding persons in the object region and for locating them in the passageway. Since the size of a person appears differently in TV images depending on the location in the passageway, we use a special map where the size of a person is invariant. We define an X–Y coordinate

(a)

(b)

Fig. 3. Extraction of a floor region. (a) Input image. (b) Floor region of image (a).

system on a plane parallel to the floor of the passageway. The origin is at the location of the TV camera and the Y axis of the system coincides with the direction of the floor of the passageway. We call this coordinate system a region map. Since the location of point P on the flat plane (shown in Fig. 1) can be calculated from the image address (Q) of P, the decline (θ) and height (H) of the TV camera and parameters of the camera model,[6] the floor area in a TV image can be transformed to the region map, as shown in Fig. 4. Simultaneously, the object areas are also mapped there. The white part in the figure represents floor area and the darker parts represent object areas.

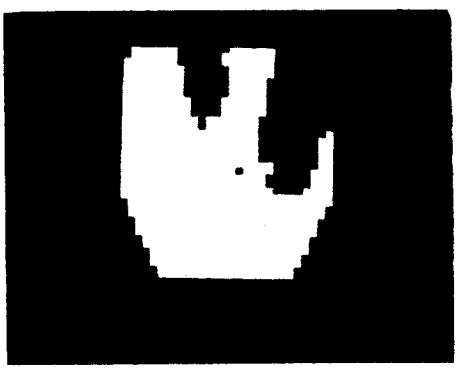

Fig. 4. Region map of Fig. 3.

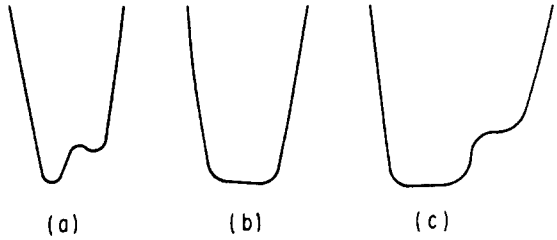

Fig. 5. Patterns of person on a region map.

We use the shapes of persons and their size as criteria for finding persons in an object area. The shapes of persons on the region map can be classified into three types, as shown in Fig. 5. Type (a) and (b) represent a person, the difference between two types being caused by the position of toes (or heels). Type (c) represents the situation where one person partly overlaps another. Figure 6 represents a simplified example of a person. Point $P1$ represents a toe which contacts the floor of the passage-way. Points $P2$ and $P3$ correspond to both sides of the legs. The width between $P2$ and $P3$ reaches a fixed value at a little distance from $P1$. Since the height of the TV camera is lower than that of a person, points $P5$ and $P6$ meet in the upper part of the region map. We decide whether an object area in the region map is a person or not using the width of and the length of the shape and its pattern.

The process is explained by using an example, as shown in Fig. 6 and Fig. 7.

(1) The region map is scanned from the bottom to the top to find the bottom point $P1$ ($X1$, $Y1$).

(2) If the upper part of the object area reaches the upper end of the region map, the contour is searched for a pair $P2$ and $P3$ such that the width $W1$ of the contour is greater than a threshold $L1$ (0.3 m) for the first time. If a pair of points $P2$ ($X2$, $Y2$) and $P3$ ($X3$, $Y2$) satisfies the condition, the person is located at the point $P4((X2 + X3)/2, Y1)$.

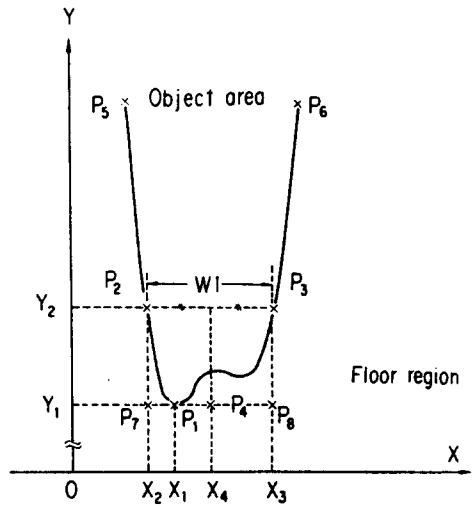

Fig. 6. Simplified example of a person on the region map.

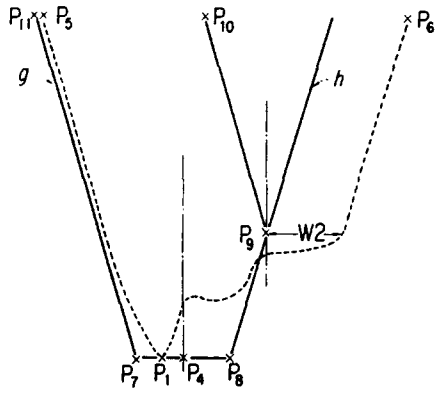

Fig. 7. Mask for finding occlusion.

(3) In order to check whether or not another person is occluded by the first person, two lines g and h, as shown in Fig. 7, are drawn from points $P7$ and $P8$, respectively. The slope of each line is equivalent to that of an outer line of the figure obtained by mapping of a cylinder with diameter $L1$ to the region map. If the width $W2$ of the protruded part under the lines is greater than a threshold $L2$(0.1 m) at point $P9$, we determine that another person is also included inside the contour. Then the slope of the line is changed to that of the line parallel to line g so that another person may be easily found later. Then only the area surrounded by points $P7$, $P8$, $P9$, $P10$ and $P11$ is deleted from the region map.

(4) This process is repeated until no person is found.

4. CORRESPONDENCE OF PERSONS

This section describes a method for finding the correspondence of persons between two consecutive region maps using location information.

We use the constraint that the maximum speed of a person is V (assumed to be 1 m/s). The distance moved by a person between two consecutive region maps is less than $V \cdot Dt$, where Dt is the time interval between two images (0.7 s). Pairs of persons are found between two region maps under this condition. If more than one candidate for a person is found, the corresponding person is determined uniquely by the following criterion. Suppose that the position of the k-th person on the first and second region maps are $(X1_k, Y1_k)$ and $(X2_k, Y2_k)$, respectively, and n is the number of positions on the first region map. The pairs are determined such that the following evaluation function may be minimized

$$E = \sum_{k=1}^{n} E_k$$

$$E_k = \sqrt{W \cdot (X2_k - X1_k)^2 + (Y2_k - Y1_k)^2} \quad W > 1.$$

E_k is a weighted distance for the k-th person. If $W = 1$, then E_k represents ordinary distance. Here W is greater than 1, because persons are assumed to move more easily along the passageway.

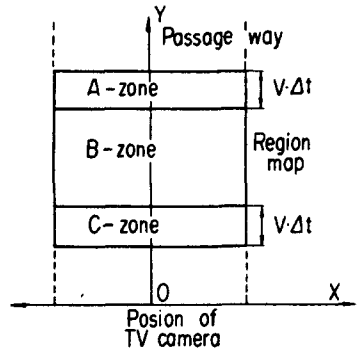

Fig. 8. Three zones of a region map.

Correspondence may not be found in the case where persons come into or out of the region map or persons are occluded by others. The former case occurs in the front region or the back region of the region map. To discriminate these regions, region maps are divided into three zones (A, B and C zones), as shown in Fig. 8. Persons are assumed to go into or out of the region map through zone A or C. The ranges of zones A and C are fixed on $V \cdot Dt$. Consequently, persons in zone B cannot go out of the region map in the time interval (Dt). First, pairs of persons are found between zone B on the first region map and the whole area on the second region map, using the criteria mentioned above. Secondly, the correspondence is found between persons in zones A and C on the first region map and the remaining persons on the second region map. If all pairs are not found, the remaining persons on both region maps are considered as follows.

The remaining persons in B zone on the first region map are determined to be behind someone on the second region map. It is also concluded that the remaining persons in zone B on the second map were occuluded by someone on the first region map and the remaining persons in zones A and C have just entered the region map from the outside area.

The remaining persons in zones A and C on the first region map are considered to have gone out of the region map.

5. EXPERIMENTAL RESULTS

The image data were taken every 0.7 s from a scene where persons were walking at ordinary speed in a passageway in our laboratory building. The floor of the passageway is made of ivory tiles and the lighting is fluorescent lamps on the ceiling. The brightness of the passage way was 200 lux on average. The height of the TV camera was 0.74 m and its viewing angle to the floor was 10°. The upper end of the region map was 6 m from the TV camera and its lower end was 2.7 m. The width of the region map was 1.4 m.

Figure 9 shows examples of four successive images. The persons in these images are located on the region map as shown in Fig. 10, and their movements are denoted by dot lines. The person at the middle in the

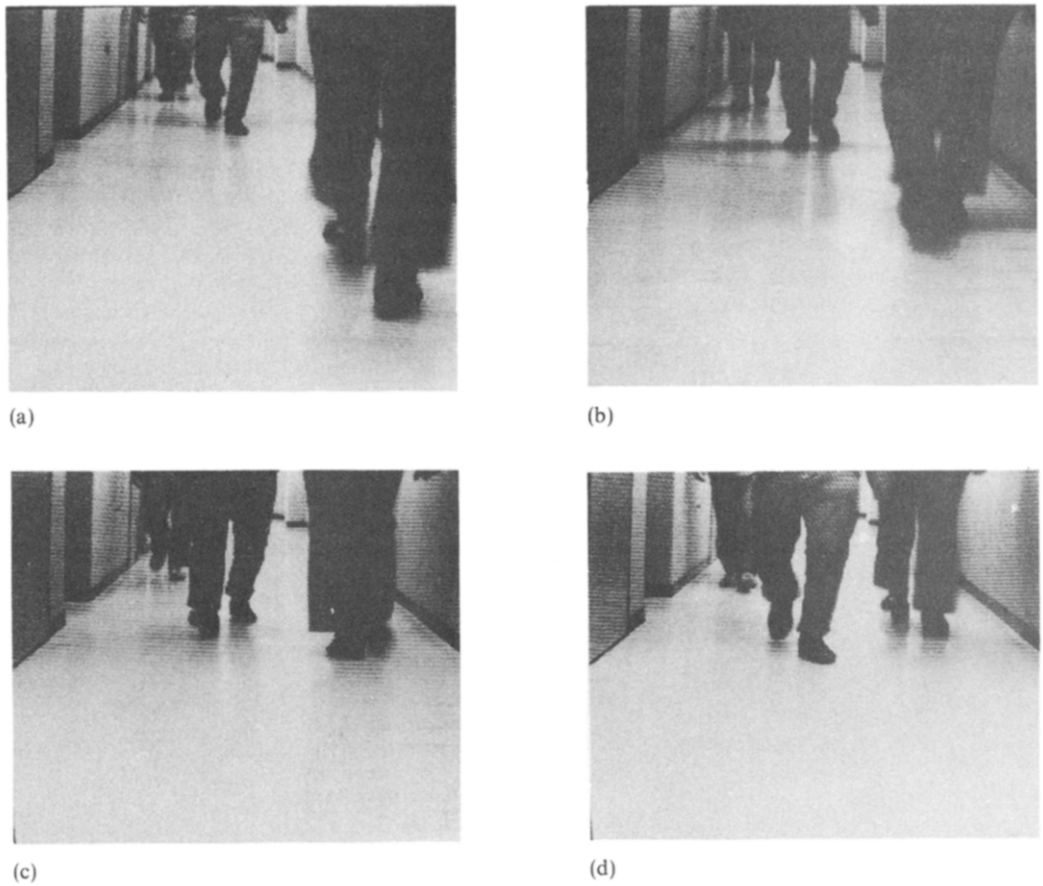

Fig. 9. Example of four consecutive images. (a) 1st scene. (b) 2nd scene. (c) 3rd scene. (d) 4th scene.

scene was out of the region map and the person at the left side also was out of the region map for a time, so they were not detected.

Figure 11(a) shows the case where a person wears a pair of white trousers. The result of image segmentation is shown in Fig. 11(b). The image processing by the mean and the variance of brightness was valid, even if persons wear white clothes similar to the floor in color.

Persons with skirts, as shown in Fig. 12(a), have bulges in the region map, as shown in Fig. 12(b). Another person seems to be occluded because of the bulge of skirt. Actually, the width of the shape is a little larger than usual but the part protruding from the two lines g and h in Fig. 7 is small.

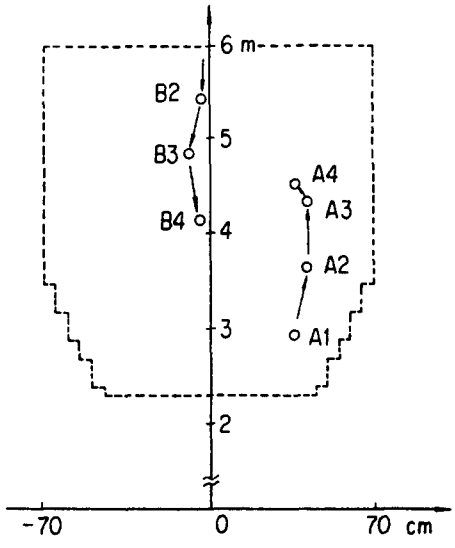

Fig. 10. Movements of persons in Fig. 9.

6. HIGH SPEED PROCESSING

The experimental system called DIP (Digital Image Processor)[7] was used for taking image data automatically and for realizing high speed image processing. The DIP hardware consists of two kinds of TV camera, seven image memories, three frame memories, display devices, a high speed image processor (PPP)[8] and an interface with PRIME-750. In the PPP basic image processing functions, such as 2-D convolution, region labeling and another logical operations for image data, can be executed at high speeds. Software utilities[9] are provided for FORTRAN users. The basic image processing, such as 2-D convolution, consumes 65 ms per image (256 × 256 pixels) and the time for region labeling is 200 ms.

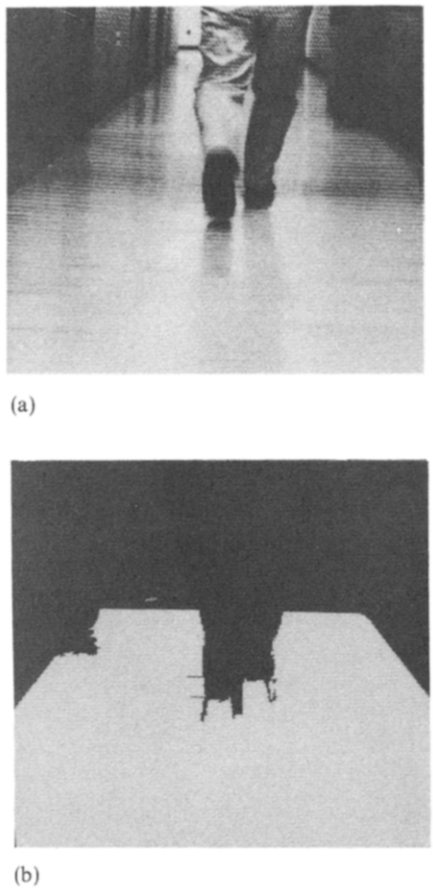

Fig. 11. Example of a person wearing white clothes. (a) Input image. (b) Extracted floor region.

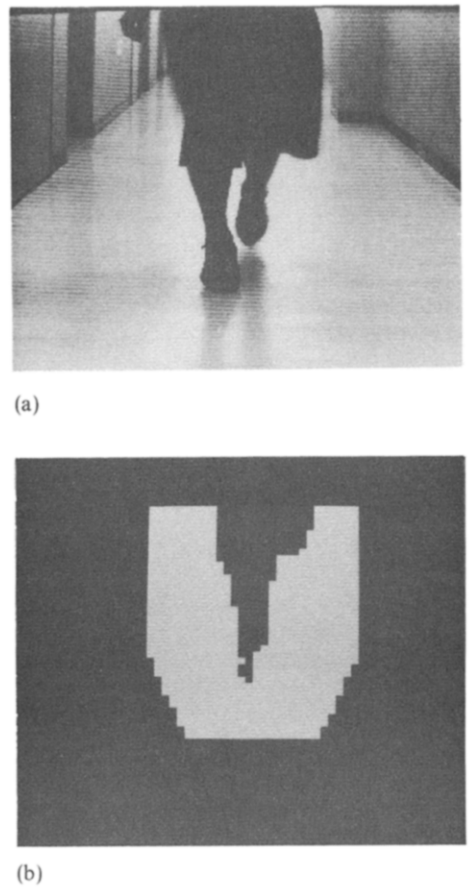

Fig. 12. Example of a lady wearing a skirt. (a) Input image. (b) Region map of the image (a).

Most operations mentioned in Sections 2 and 3 can be processed in the PPP, and the other operations are processed in the host computer. For example, each calculation of the equations $M(i, j)$ and $D(i, j)$ can be made using 2-D convolution and absolute operation and add operation of the PPP. To delete small areas surrounded with floor areas, mentioned in Section 2, region labeling and histogram function were used. It takes 2 s to find the floor region by the PPP, including transfer of data between processors and memories. It takes about 5 s to map the floor region onto the region map and to find the correspondence of the persons by the host computer. However, the development of hardware, such as a concurrent multiprocessor system, could make the total time smaller than 0.7 s.

7. CONCLUSION

We have proposed a method for finding the movements of persons in a sparse sequence of TV images. Persons were extracted from each image and their actual locations in the passageway were calculated by inverse perspective transformation. The movements of persons were found by comparing the location changes between two consecutive images.

The occlusion problem was solved by predicting approximate locations of persons at every image based on their assumed maximum speed. The possibility of real time processing was checked using our experimental system, called DIP.

Detection of the movements of persons might be more reliable and efficient if the motion directions of persons obtained from previous images are used, as well as location information on persons.

REFERENCES

1. H. H. Nagel, Overview on image sequence analysis, *Proceedings of Image Sequence Processing and Dynamic Scene Analysis*, pp. 2–39. Springer-Verlag (1983).
2. L. Dreschor and H. H. Nagel, On the frame-to-frame correspondence between greyvalue characteristics in the images of moving objects, *German Workshop on Artificial Intelligence*, pp. 18–21. Springer-Verlag (1981).
3. C. L. Fennema and W. B. Thompson, Velocity determination in scenes containing several moving objects, *Comput. Graphics Image Process.* **9**, 301–315 (1979).
4. B. K. P. Horn and B. G. Schunk, Determining optical flow, *Proceedings of Image Understanding Workshop (DARPA)*, pp. 144–156 (1981).
5. R. Jain and H. H. Nagel, On the analysis of accumulative difference pictures from image sequences of real world scenes, *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI 1**, 206–214 (1979).
6. R. O. Duda and P. E. Hart, *Pattern Classification and*

Scene Analysis, pp. 379–404. Wiley-Interscience (1973).
7. T. Tsukiyama and M. Suwa, Digital image processor—an experimental tool for image processing research, *Bull. electrotech. Lab., Japan* **44**, 606–614 (1980) (in Japanese).
8. K. Mori and H. Asada, Design of Parallel Pattern Processor for Image Processing, *Proceedings of National Computer Conference*, pp. 1027–1031 (1978).
9. T. Tsukiyama and Y. Shirai, A software system for the digital image processor, *Bull. electrotech. Lab., Japan* **46**, 439–461 (1982) (in Japanese).

About the Author—TOSHIFUMI TSUKIYAMA was born in Nara, Japan, on 27 June 1948. He received a B.E. degree in Biophysical Engineering from Osaka University in 1972. He joined the staff of the Electrotechnical Laboratory of the Japanese Government in 1972. He is currently a senior member of the Computer Vision Section. His research interests include robot vision, motion estimation and architecture for image processing. Mr. Tsukiyama is a member of the Information Processing Society of Japan and the Society of Instrument and Control Engineers of Japan.

About the Author—YOSHIAKI SHIRAI was born in Toyota, Japan, on 3 August 1941. He received a B.E. degree from Nagoya University in 1964 and a Ph.D. degree in Mechanical Engineering from Tokyo University, Tokyo, Japan, in 1969.

In 1969 he joined the staff of the Electrotechnical Laboratory, where he conducted research in recognition of objects in the PIPS project. He is currently Chief of the Computer Vision Section. From 1971 to 1972 he was a Visiting Researcher in the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, where he worked on computer vision. He received the Pattern Recognition Society Award, and an Honorable Mention for the Pattern Recognition Society Award in 1972 and 1979, respectively. He is the author of the book *Computer Vision* (in Japanese). His research interests are computer vision, medical image processing and artificial intelligence.

Dr. Shirai is a member of the Institute of Electronics and Communication Engineers of Japan and the Information Processing Society of Japan.