A COMPREHENSIVE SOLUTION TO COMPUTER VISION BASED GROUP SUPERVISORY CONTROL

Albert T.P. So, W.L. Chan, City Polytechnic of Hong Kong
H.S. Kuok, S.K. Liu, Chevalier (HK) Ltd.

ABSTRACT

A paper presented at ELEVCON'92 entitled "A Computer Vision Based Group Supervisory Control System" suggested that the application of computer vision improve both the quantity and quality of service of conventional group supervisory control systems. However, such fuzzy logic based primitive design limited the positioning of the cameras due to the nature of the image interpretation algorithm. For economic reasons, it has been suggested that security cameras installed at the corners of lift lobbies should be used. Therefore, two new methods, based on "Algebraic Reconstruction Technique" and "Depth From Motion", have been developed and presented in this paper, that can upgrade the capability of the existing system to estimate the number of passengers in the lobbies more accurately.

1. REVIEW OF THE EXISTING COMPUTER VISION BASED SUPERVISORY CONTROL

It has been suggested [1] that perfect control can be achieved if the group supervisory control of a lift system can actually know every detail of the traffic flow. The information includes the number of passengers waiting at the lobby, their destinations as well as the number of passengers inside each lift car. The solution to the second item, i.e. passenger's destinations, has been suggested by the ACA's optimal computer group control [2] while the first and third items are handled by computer vision [1] and other less precise but perhaps faster and inexpensive methods [3-4]. By employing computer vision, it is possible to estimate the exact number of passengers inside a lift car as well as waiting at the lift lobbies under a real-time basis.

1.1 Imperfections in conventional group control

The conventional car allocation algorithm has a lot of imperfections. It only considers two aspects, namely the distance between the current position of a lift car and the landing call demand floor and the number of foreseeable stoppages when the lift car travels from the current floor to the landing call demand floor. Certain factors have not been considered, including the number of passengers initiating such a landing call, the ratio of up-direction and down-direction passengers, the spare capacity inside the lift car and the space/weight ratio of passengers inside the lift car etc. All these factors seriously affect the efficiency of car allocation procedures in various ways.

1.2 Improvements by computer vision

The following improvements can be achieved if the number of passengers are known at appropriate locations.

a) Landing calls initiated by a large group of passengers can be given priority automatically.

b) Any floor occupied by a large group of passengers can be assigned as a preferential floor or heavy demand floor automatically.

c) A lift car with over 90% load can serve one or two more passengers.

d) Any landing floor with landing button on but without any passenger can be ignored automatically.

e) Holding time for car doors becomes unnecessary.

f) Up-direction passengers and down-direction passengers can be discriminated automatically.

g) Up-peak and down-peak control modes can be switched in and out automatically.

h) Other minor improvements such as defeating the anti-nuisance function when a single child of light weight is inside the lift car.

1.3 Imperfection with the existing computer vision based supervisory control system

The existing fuzzy logic based passenger number estimation algorithm [1] restricts the location of the camera. It should always be placed right up in the ceiling of the lift car or the landing floor so that the top view of the whole area of interest is covered. Only by this arrangement that the passengers appear as individual patches on the image where, most importantly, the area of the patch has a direct relationship with the number of passengers corresponding to the patch. This approach is reasonably satisfactory for applications in the lift cars but certainly there are problems when it is used for the lift lobbies. First of all, the common ceiling height of most lift lobbies is no higher than four to five metres and thus a standard optical lens cannot cover the whole lobby unless a wide-angle lens is employed but it will heavily distort the images. Secondly, it is expensive for both installation and maintenance if the camera is to be located at the middle of the ceiling. Also, for economic reasons, the security cameras located at the corners should be used instead. In this case, the existing fuzzy logic based passenger number estimation algorithm cannot be applicable anymore since the area of the patch is no longer proportional to the number of passengers belonging to the patch. Two new approaches are suggested here.

2. THE "ALGEBRAIC RECONSTRUCTION TECHNIQUE" APPROACH (ART)

The basic idea of this method comes from the operating principle of a piece of standard medical imaging equipment, namely X-ray Computer Tomography, or CT-scanner. The main application of CT-scanners is the diagnosis of abnormalities within the skull. Head scanning is used to detect causes of neurological disorders such as brain neoplasms, infarctions, cerebral edema, abscesses and ophthalmologic diseases etc. The idea of CT comes from the need for a technique aimed at computing true sectional views from projection data. Reconstruction of an object from its projections is a problem of linear algebra that can be solved in a straightforward manner by matrix inversion. For our application, two to three cameras are installed at the corners and they are focused towards the lobby. Each camera produces an image of the lobby as viewed form its direction of focus. Image subtraction between the real time image obtained and a stored reference image of the vacant lobby results in patches of passengers. Each patch may belong to a few passengers when they are overlapping one another as viewed from the camera. Our job is to reconstruct the distribution of passengers in the lobby by combining the two to three images associated with the respective cameras so that the number of passengers in the lobby can be estimated. Before we look at the system, it is preferable to have a brief understanding of the ART algorithm.

2.1 Principle of ART [5]

There are basically two approaches, namely the iterative approaches and the direct reconstruction method which involves the complicated Fourier Transform. Although the modern CT-scanners are employing the direct approach, we are adopting the iterative approach due to its simplicity and light computational burden. The mathematics involved is a relatively old, but seldomly used, field of study involving the reconstruction of a two-dimensional distribution from its projections. The most straightforward, although computationally inefficient solution involves linear algebra. The two-dimensional image is reconstructed using a matrix inversion of the projection data. For images of reasonable complexity, the attempt is to find a two-dimensional distribution that matches all of the projections. An initial distribution is assumed and it is compared with the measured projections. Using one of a variety of iterative algorithms, the initial distribution is successively modified. Reference is made to Fig. 1 where one line is highlighted.

The ART is based on the very general premise that the resultant reconstruction should match the measured projections. The iterative process is started with all reconstruction elements f_i set to a constant such as zero or the mean f_{jm} for the jth line of projection. f_{jm} is given by

$$f_{jm} = \frac{g_j}{N} \tag{1}$$

In each iteration, the difference between the measured data for a projection g_j and the sum of the reconstructed elements along that ray, f_{js}, is calculated where

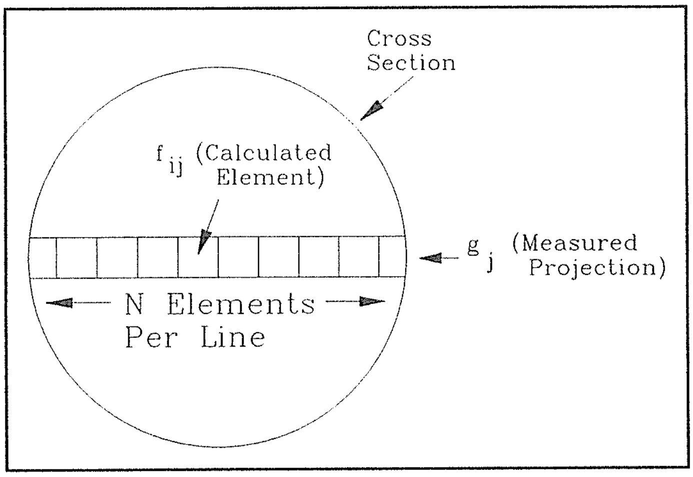

Fig. 1

$$f_{js} = \sum_{i=1}^{N} f_{ij} \tag{2}$$

Here, f_{ij} represents an element along the jth line forming the projection ray g_j. This difference is then evenly divided among the N reconstruction elements. The iterative algorithm is defined as

$$f_{ij}^{q+1} = f_{ij}^{q} - \frac{g_j - \sum_{i=1}^{N} f_{ij}^{q}}{N} \tag{3}$$

where the superscript q indicates the current number of iteration step. The algorithm recursively relates the values of the elements to those of the previous iteration. A number of variations on this general theme have been proposed. One nonlinear formulation makes use of the known non-negativity of the density values f_{ij}. Thus, where $f_{ij} < 0$, it is set equal to zero. Another variation is known as multiplicative ART, as compared to the previous original algorithm, which is additive ART. In the multiplicative version, the original density values are multiplied by the ratio of the measured line integral g_j to the calculated sum of the reconstructed elements. This is given by

$$f_{i\,j}^{q+1} = \frac{g_j}{\sum_{i=1}^{N} f_{i\,j}^{q}} f_{i\,j}^{q} \qquad (4)$$

In multiplicative ART, each reconstructed element is changed in proportion to its magnitude. This is in sharp contrast to additive ART, where each element in the ray is changed a fixed amount, independent of its magnitude.

2.2 Application of ART in computer vision based supervisory control

For illustrative purpose, two cameras (A and B) are installed at the corners of the lobby, as shown in Fig. 2. The lift lobby under interest is enclosed by the small square at the middle. The image from each camera is segmentated into various vertical stripes or columns so that the lift lobby is divided into various zones from each camera point of view, six in this example. Passengers (P1, P2 and P3) are overlapping one another as seen by camera B and they occupy zone column B_3 of camera B's image. However, they are separated from one another as seen by camera A, occupying zones A_2, A_3 and A_4 of camera A's image. Passenger P5 appears as an individual object as seen by both cameras, occupying Zone A_5 and B_5. Passenger P4 is also an individual object occupying zone A_2 but only half of it occupies zone B_4 and the remaining half occupies zone B_5. In this way, overlapping of passengers may lead to false information to one camera but this situation can be partly compensated by another camera, thus minimising the total error obtained. Image subtraction between a real-time image and a reference image for the vacant lobby can extract all passengers in the form of patches. Experience has revealed that a non-reflective interior surface design for the lobby installed with highly directional artificial lighting can give the best

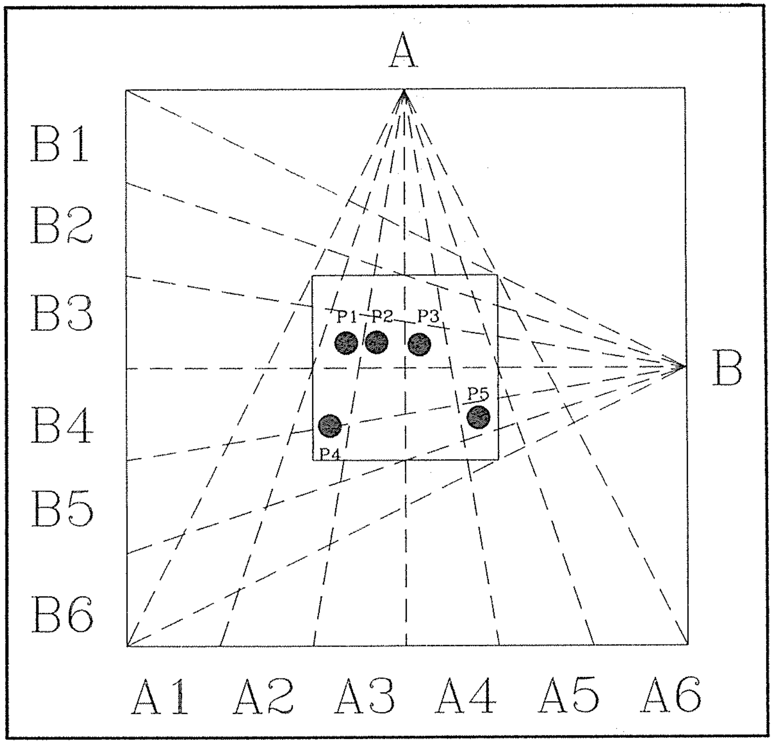

Fig. 2

for patch enhancement. For each image, we therefore have six vertical white columns superimposed by black patches. A degree of occupancy, μ, of each column, in terms of percentage of patches in the associated column, is calculated. The area of patches or the length of patches can be used for identifying this degree of occupancy. Since a passenger may appear as a small patch if he/she is standing at a remote end from the camera, correction with respect to distance from camera is carried out. Each column can further be divided into sub-zones. The degree of occupancy of each sub-zone is corrected in accordance with the physical distance of the real scene of the sub-zone from the camera and the degree of occupancy of the whole column is the addition of all the degrees of all the sub-zones. After image grabbing, image subtraction, thresholding and distance correction, $\mu(A_k)$ for $k = 1,..,6$ and $\mu(B_k)$ for $k = 1,..,6)$ are known. By AST, $\mu(A_iB_j)$ for $i,j \in \{1,2,..,6\}$ can be estimated. The meaning of $\mu(A_iB_j)$ is actually the degree of occupancy of the zone in the lobby generated by the intersection of zone A_i and B_j. For example, Passenger P1 is in fact occupying the intersection zone of zone A_2 and B_3. We shall expect a high value for $\mu(A_2B_3)$ whereas $\mu(A_4B_5)$ should have a value approaching zero. Integrating all the μ's over the whole lift lobby and multiplying the resultant value by the maximum number of passengers allowed for the whole lobby, we can get the real-time number of passengers at the lobby. Also, by checking the distribution of degree of occupancy over the lobby, it is possible to tell where are the passengers grouped together.

2.3 Comments on the AST approach

This method is a fast and reasonably accurate method of estimating the number of passengers at the lift lobby. Iteration only takes seconds to complete and only one image from each camera is deemed sufficient to generate the result. Time multiplexing can conveniently be implemented so that one image grabber card is adequate to be interfaced to tenths of optical cameras. However, errors caused by overlapping of passengers on 2-dimensional images still exist although the installation of more cameras can reduce such error. It is professionally felt that three cameras for each lift lobby is an optimal arrangement.

3. THE "OPTICAL FLOW" APPROACH

In order to overcome the problem of overlapping, a "field of depth" of the lobby scene is to be generated. Provided such field is compiled, two passengers can be discriminated between one another even though one forms an obstacle to the other along the line of sight of the camera. This is in fact in line with human reasoning. Two cameras are required, place side by side as if human eyes do. They are first of all well calibrated and two images of the same scene are grabbed. Correspondence between pixels is achieved by generating a velocity profile. Once correspondence is set up, the depth of the scene point associated with the particular pixel can be found. Clustering all the scene points can indicate the number of passengers in the lift lobby.

3.1 Camera calibration

The two CCD optical cameras (a and b) forming a stereoscopic system are placed side by side and well calibrated in accordance with Tsai's method [6]. After the calibration, every point in the world coordinate (x_w, y_w, z_w) can be transformed into the two camera-orientated coordinate systems (x_a, y_a, z_a) and (x_b, y_b, z_b) and further more, the frame memory coordinate systems (X_a, Y_a) and (X_b, Y_b) respectively by the following equations:

$$\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} = \begin{bmatrix} r_{1i} & r_{2i} & r_{3i} \\ r_{4i} & r_{5i} & r_{6i} \\ r_{7i} & r_{8i} & r_{9i} \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} + \begin{bmatrix} T_{xi} \\ T_{yi} \\ T_{zi} \end{bmatrix}$$

$$\frac{d_{xi}\, X_i}{s_{xi}} \left(1 + k_{1i}\rho_i^2 \right) = f_i \, \frac{r_{1i}\, x_w + r_{2i}\, y_w + r_{3i}\, z_w + T_{xi}}{r_{7i}\, x_w + r_{8i}\, y_w + r_{9i}\, z_w + T_{zi}} \tag{5}$$

$$d_{yi}\, Y_i \left(1 + k_{1i}\rho_i^2 \right) = f_i \, \frac{r_{4i}\, x_w + r_{5i}\, y_w + r_{6i}\, z_w + T_{yi}}{r_{7i}\, x_w + r_{8i}\, y_w + r_{9i}\, z_w + T_{zi}}$$

$$where \quad \rho_i = \sqrt{\left(d_{xi} \frac{X_i}{s_{xi}} \right)^2 + (d_{yi}\, Y_i)^2}$$

where i can represent either camera (a) or camera (b). The nine r's represent the rotational matrix while the three T's represent the translational matrix. f_i is the focal length; k_{1i} and s_{xi} are intrinsic parameters representing the distortion with the lens and CCD array; d_{xi} and d_{yi} are the physical separation between two neighbouring light sensors on the CCD array. When all these are known, the two cameras have been successfully calibrated.

3.2 Velocity field computation

Two images of the scene are taken for each camera with $\delta t = 0.5$ second apart. A gradient constraint equation is introduced that relates velocity of a pixel on the image, i.e. $(u_{x,y}, v_{x,y})$ to the image brightness function P(x,y,t). It can be assumed that:

$$\begin{aligned} P(x, y, t) &= P(x+\delta x, y+\delta y, t+\delta t) \\ Hence, \quad 0 &= P_x\, u + P_y\, v + P_t \\ where \quad P_x &= \frac{\partial P}{\partial x}, \; P_y = \frac{\partial P}{\partial y}, \; P_t = \frac{\partial P}{\partial t} \\ u &= \frac{\partial x}{\partial t} \Big|_{x,y}, \quad v = \frac{\partial y}{\partial t} \Big|_{x,y} \end{aligned} \tag{6}$$

The gradient constraint equation has no solution by itself and therefore an error optimisation, E, has to be used [7] such that

$$E^2 = \int\!\!\int \left\{ [P_x u + P_y v + P_t]^2 + k^2 \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 \right] \right\} dx\, dy \tag{7}$$

where k controls the relative cost of deviations from smoothness and deviations from the motion constraint and it is usually set to 1. The equation set can be solved by discrete approximation and then numerical iteration, resulting in the following expressions for the nth iteration.

$$u^{(n+1)} = \bar{u}^{(n)} - P_x \left[\frac{P_x \bar{u}^{(n)} + P_y \bar{v}^{(n)} + P_t}{3k^2 + P_x^2 + P_y^2} \right]$$

$$v^{(n+1)} = \bar{v}^{(n)} - P_y \left[\frac{P_x \bar{u}^{(n)} + P_y \bar{v}^{(n)} + P_t}{3k^2 + P_x^2 + P_y^2} \right] \qquad (8)$$

$$where \quad \bar{u} = \frac{\left(\sum_{i=-1}^{1} \sum_{j=-1}^{1} u_{x+i\,,\,y+j} \right) - u_{x\,,\,y}}{8} \quad , \quad \bar{v} = \frac{\left(\sum_{i=-1}^{1} \sum_{j=-1}^{1} v_{x+i\,,\,y+j} \right) - v_{x\,,\,y}}{8}$$

Iteration is considered a completion when $| u^{(n+1)} - u^{(n)} | + | v^{(n+1)} - v^{(n)} |$ is smaller than a threshold value.

3.3 Pixel correspondence

Image subtraction between a real time image and the reference vacant lobby image can highlight the patches due to passengers. We are only interested in the edges of the patches and thus the velocity field computation concentrates on the edge pixels only. The job at present is to find out pixel-pairs on images from the two cameras that are corresponding to the same spot in space. Provided that such pixel-pairs are located, the coordinates of the scene spot in space can be calculated and hence, its distance from the stereoscopic camera system. Each relevant pixel (X_o, Y_o) on the patch of an image from camera (a) corresponds to a straight line in space passing through the focal point of the lens. The coordinates of each point on this line in the world coordinate system can be represented by [$x_{wo}(z_{wo})$, $y_{wo}(z_{wo})$, z_{wo}], which is an equation with z_{wo} as the running parameter:

$$\begin{bmatrix} AX_oT_{za} - T_{xa} \\ BY_oT_{za} - T_{ya} \end{bmatrix} = \begin{bmatrix} f_ar_{1a}-r_{7a}AX_o & f_ar_{2a}-r_{8a}AX_o & f_ar_{3a}-r_{9a}AX_o \\ f_ar_{4a}-r_{7a}BY_o & f_ar_{5a}-r_{8a}BY_o & f_ar_{6a}-r_{9a}BY_o \end{bmatrix} \begin{bmatrix} x_{wo} \\ y_{wo} \\ z_{wo} \end{bmatrix} \qquad (9)$$

$$where \quad \begin{cases} A = \dfrac{d_{xa}}{s_{xa}} (1 + k_{1a} \rho_a^2) \\ B = d_{ya} (1 + k_{1a} \rho_a^2) \end{cases}$$

This line can be mapped onto the image from camera (b) by using equation set (5) to form an epi-polar line. This epi-polar line will intersect the edges of the patches at various points. The velocity profile at these points are checked with the velocity profile of (X_o, Y_o) . The pixel with a velocity vector more or less identical to that of (X_o, Y_o) is chosen as correspondence. With two pixels on the two images from two cameras, it is possible to fix a point in the world coordinate system.

3.4 Scene spots fuzzy clustering

After the completion of step 3.3, we shall have a set of points in the world coordinates. These points are marked on a horizontal plane, i.e. the z-coordinate is being ignored. These points are actually on the external surfaces of the passengers seen by the cameras. The effect of fuzzy clustering [8] is to identify whether a group of points belongs to one or more passengers. Assume that there are n number of points to be clustered into a number of groups. The set of points is defined as

$$U = \{ x_1 , \ldots , x_n \} \subset \Re^2 \qquad (10)$$

$$\textit{where} \quad x_i = (x_i , y_i)^T \; : \; \textit{coordinates of ith point}$$

A relational matrix S_{nxn} is set up below as

$$S_{n\,x\,n} = \begin{bmatrix} r_{11} & \cdot \; \cdot & r_{1n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ r_{n1} & \cdot \; \cdot & r_{nn} \end{bmatrix} \qquad (11)$$

where $0 \leq r_{ij} \leq 1$: i,j = 1,...,n are the degree of correlation between point i and point j and it can be evaluated from the characteristics values by a number of ways. In our case, since we are using the points to distinguish between passengers, the size of a standard passenger is a natural choice of the degree of correlation. d_o, the normal diameter of a human being from a top view, which is around 0.4 m, is chosen as a datum. Therefore, r_{ij} can be defined below as:

$$\begin{aligned} r_{ij} &= 1 - C\, d\,(x_i , x_j) \\ \textit{where} \quad d\,(x_i , x_j) &= | x_i - x_j | \\ &= \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \qquad (12) \\ C &= \begin{cases} 0 & \textit{if } d\,(x_i , x_j) \geq d_o \\ \dfrac{1}{d_o} & \textit{if } d\,(x_i , x_j) < d_o \end{cases} \end{aligned}$$

An equivalence relational matrix can be worked out by:

$$(S_{nxn})^N = S_{nxn} \, o \, (S_{nxn})^{N-1} = (S_{nxn})^{N+1} = (S_{nxn})^{N+2} = \cdot\,\cdot \qquad (13)$$

such that the operation "o" is defined as follows:

$$\begin{aligned} R \, o \, Q &= [r_{ij}] \, o \, [q_{ij}] = P = [p_{ij}] \\ &= \Big[\sup_{k = 1 \text{ to } n} \big(\textit{inf} \, (r_{ik} , q_{kj}) \big) \Big] \end{aligned} \qquad (14)$$

It can be shown that any equivalence relationship matrix defines a unique partition of a set. By α-cutting the equivalence relational matrix, crisp clustering is achieved. Those elements, $r_{ij} = 1$, imply point i and point j belong to the same group. By checking against the size of a group, it is possible to estimate the number of passengers seen by the two cameras.

4. CONCLUSIONS

The existing computer vision based group supervisory control system has been briefly reviewed. Two new approaches have been introduced in this paper. The AST approach is fast and economical but its accuracy is not so high. The accuracy can be improved by installing more cameras and increasing the ceiling height so that the level of overlapping of passengers can be minimised. The "Optical Flow" approach is more accurate but it is very computational intensive. It is based on one critical assumption that only moving passengers can be detected. Stationery objects in the lift lobby can not be identified since the velocity vectors of edges of these objects have very small amplitudes. However, both methods have a common merit. They need a limited number of cameras which have been installed at the lift lobbies for security purposes and they need only one or two images to arrive at the solution. Hence, the hardware cost is low in general. Further research work is deemed necessary to emulate the real visual cognitive ability of the high-level biological visual system of human beings.

5. ACKNOWLEDGEMENTS

This research project is supported by the Strategic Research Grant #700186 of City Polytechnic of Hong Kong. The research team would also like to thank IAEE (Hong Kong Branch) for the support and advice.

6. REFERENCES

[1] So A.T.P., Chan W.L., Kuok H.S. and Liu S.K. "A Computer Vision Based Group Supervisory Control System", in Barney G.C. Eds., Elevator Technology 4, 1992, pp. 249-258.

[2] Barney G.C. Eds. "Elevator Traffic Analysis, Design and Control", P. Peregrinus, 1985.

[3] Siikonen M.L. "Simulation - A Tool for Enhanced Elevator Bank Design", Kone Elevators Research Centre, Elevator World, April, 1991.

[4] Al-Sharif L.R. and Barney G.C. "The Inverse S-P Method Deriving Lift Traffic Patterns from Monitored Data", Control Systems Centre Report, No. 745, August, 1991.

[5] Macovski A. Eds. "Medical Imaging Systems", Prentice-Hall Inc., N.J., 1983, pp. 114-117.

[6] Tsai R.Y. "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses", IEEE Journal of Robotics and Automation, Vol. RA-3, No. 4, August, 1987, pp. 323-344.

[7] Horn B.K. and Schunck B.G. "Determining Optical Flow", Artificial Intelligence, Vol. 17, 1981, reprinted in Computer Vision, Brady M., Eds., Amsterdam, pp. 185-203.

[8] Zimmermann H.J. Eds. "Fuzzy Set Theory and Its Applications", 2nd Ed. Kluwer Academic Publishers, 1990, pp. 220-240.