*IEE Electronic and Communications Engineering Journal, Vol. 7, No. 1 (Feb), pp.37-47*

# Crowd Monitoring Using Image Processing

The understanding of crowd behaviour in semi-confined spaces is an important part of the design of new pedestrian facilities, for major layout modifications to existing areas and for the daily management of sites subject to crowd traffic. Conventional manual measurement techniques are not suitable for comprehensive data collection of patterns of site occupation and movement. Real-time monitoring is tedious and tiring, but safety-critical. This article presents some image processing techniques which, using existing closed-circuit television systems, can support both data collection and on-line monitoring of crowds. The application of these methods could lead to a better understanding of crowd behaviour, improved design of the built environment and increased pedestrian safety.

**by Anthony C. Davies, Jia Hong Yin and Sergio A. Velastin**

Department of Electronic and Electrical Engineering, King's College London, United Kingdom

### Introduction

Although crowds are made up of independent individuals, each with their own objectives and behaviour patterns, the behaviour of crowds is widely understood to have collective characteristics which can be described in general terms. For example, descriptions such as 'an angry crowd', a 'peaceful crowd' etc. are well accepted. The use of terms such as “mob”, “mob rule”, have at least since Roman times, carried an implication that a crowd is something other than the sum of its individual parts and that it may possess behaviour patterns which differ from the behaviour expected individually from its participants.

It might be assumed that if there is some scientific basis for the study of crowd behaviour, it must belong properly within the social sciences and psychology, and that the physical sciences and engineering have no business in getting involved with such studies. However, electronic engineers have an understanding of field theory and of flow dynamics which may provide insight into characteristics of crowd behaviour, and can also provide the expertise to suggest solutions to crowd monitoring and control based on technological developments in image processing and image understanding.

In the search for a scientific understanding of crowd behaviour there is a partial analogy with achievements in the study of gases. A gas is made up of individual molecules, moving about more-or-less independently of one another, with differing velocities and directions. However, the ideal-gas theory provides an extremely accurate basis for predicting the properties and behaviour of gases over a wide range of conditions, without taking any detailed consideration of the behaviour of individual molecules. This theory represents a major scientific achievement which could hardly have occurred if the notion had prevailed that the equations of motion for each individual molecule have to be solved in order to predict the overall behaviour of a gas in any particular situation.

Of course, the individuals making up a crowd are much more variable and complex than the molecules making up a gas, but it is at least reasonable to hope that some mathematical rules of behaviour might be derived which could usefully approximate the behaviour of real crowds. The observation of commuters emerging from a busy station concourse during a morning rush-hour, and the manner in which such a crowd flows past fixed obstacles, must remind electrical engineers of the behaviour of charged particles moving under the influence of an electrical field. As the crowd density and velocity increase, changes can be observed from what may resemble laminar flow to something having some characteristics of turbulence, providing an analogy with fluid flow in aerodynamics. This reinforces the common-sense idea that major obstacles should be designed to have smoothly rounded corners and suggests that in the design of such concourses something could be learned from the flow of air around an aerofoil. If some individuals are moving in the opposite direction (like particles of the opposite charge) collisions may be observed occasionally if the crowds are moving too fast or if some individuals have too much inertia (for example, individuals pushing overloaded baggage trolleys).

## Managing and Understanding Crowds

Particularly in recent years, television news reporting has presented images of huge crowds demonstrating peacefully for political change, while also bringing to our attention the dangers of large crowds unable to achieve their collective objectives (for example, resulting in disasters at football matches) and dangers arising from failures in controlling their movements in an orderly way when trying to evacuate areas subject to terrorist threats or fire.

The management of crowds at football matches, pop concerts, carnivals, airports and even in the day-to-day movement of commuters in and out of large cities is a substantial problem with serious consequences for human life and safety and for public order if it is not managed successfully.

There are two aspects to this problem to which a scientific understanding of crowd behaviour could contribute. One is in the design of the environments in which crowds are expected to arise and the other is in the real-time monitoring and control of crowds within existing, typically urban, structures. There is, of course, some need to manage crowds on open areas (e.g. refugees congregating in rural regions) but the lack of a built environment of constraining buildings make the problems rather different and more related to the distribution of food and medical assistance.

*Designing or Modifying the Built Environment*

The development of models of crowd behaviour would provide a basis for informing architects and town-planners. Such models could also supplement existing methods of prediction through their use in computer-simulation to investigate the characteristics of crowd flow through proposed designs. It is well known that crowd flow through airport and station concourses is related to the line-of-sight topology [1] of the structure. If there are many obstacles preventing individuals seeing their intended exit or other destination, they may not take to optimum route and may move more slowly. A larger crowd-storage capacity must then be provided to prevent dangerous levels of congestion, with clear implications for the evacuation rate that would be possible (e.g. in conditions of electrical power failure, darkness, and perhaps thick smoke). The free flow of crowds through such man-made facilities is likely to be enhanced by long line-of-sight paths, smoothly contoured obstacles and rounded rather than angular corners in corridors, and a segregation of high-density simultaneous flows in opposing directions. Poorly located exits and entrances from a large open area can give rise to rotational movements in crowds which hinder free-flow and can be dangerous in conditions near to maximum capacity.

All this is common-sense and can be taken into account without any quantitative study. However, if aspects of such crowd behaviour could be expressed in the form of mathematical laws, this would usefully supplement heuristic and experience-based design of such facilities, and as mentioned above, would provide a foundation for better computer simulation and hence prediction of the suitability of planned facilities.

When space is at a premium (as in most large cities), architects may wish to seek optimum solutions, for which the support of mathematically-based theory could be very helpful.

*Monitoring Crowds for Safety*

For facilities already in existence, there is an established practice of using extensive closed-circuit television monitoring of crowds. This is used in a number of ways. For example, the television monitors are routinely observed to look for problem conditions. These may be problems arising from various forms of congestion or problems arising from individual incidents not related to a crowd, such as a theft, the outbreak of a fire, or the placing of a bomb by a terrorist. Of course, the cameras installed primarily in response to the latter class of problem (e.g. for what may be collectively be termed 'security') are then available for use for the former class of problems (e.g. for what may be collectively be termed 'crowd safety'). When congestion (crowd density) exceeds a certain level (this level being dependent upon the collective objective of the crowd and the environment), danger may occur for a variety of reasons. Physical pressure may result directly in injury to individuals or to the collapse of parts of the physical environment (for example the collapse of barriers). Alternatively the mood of the crowd may change for the worse if they sense frustration in achieving their objective, whether this objective is to get into a football stadium, to fight with the supporters of the opposing team, or to get home at the end of the working day when some emergency has shut down the commuter transport system throughout a large metropolitan region.

It is normal practice in the management of such crowds to control the arrival rate, stopping it completely when the crowd within a particular area has become too great for safety. This applies both to the entry to an auditorium or a football stadium when it is judged to be full to capacity, as well as to entry to a railway or subway station where there is a continual flow due to passengers leaving on departing trains, and where the entry does not need to be permanently stopped. In the latter case, input flow may either be periodically stopped while the internal crowd is reduced by departures or else may be slowed down by reducing temporarily the number or width of entry-doors. Of course, all such approaches lead to the probability of a new crowd congestion problem being created in a different section, formed by those individuals who have been denied access to their target area.

These requirements have led to a continuing increase in the number of installed video cameras, so that more and more public areas are subject to surveillance. In some cases, the systems include continual video recordings, usually by some form of time-lapse technique to make the storage requirement economically viable. However, the successful use of these facilities presents some problems. Human observers are normally positioned to watch the TV monitors, and it is often not considered cost-effective to have one monitor per camera and it is even less likely to have one observer per monitor.

Therefore a typical observer is confronted with a two dimensional array of monitors, a switch to select the one or ones from which a time-lapse video recording is made and, in some cases, means for a pan and zoom control of various individual cameras. In the case of short-term events (such a pop-concerts and football matches) it is reasonable to assume a high level of concentration by the observers as well as a high level of experience. However, in the routine monitoring at airports, shopping areas and station concourses, the observers are likely to lose concentration, and since significant events are very infrequent, these may either not be observed at all, or else only when it is too late to take effective action. This situation is aggravated in some cases by management noticing that the observers are not active for the majority of the time, and allocating to them other duties such as routine form-filling tasks, which reduces the rigour of their observations even more. The observers may also be tempted to read newspapers or play cards to relieve their boredom if they are not themselves adequately monitored.

### *The need for Automation*

It is clear, therefore, that there would be a substantial advantage if some form of image processing could be used on the video images, to automatically spot crowd problems as they arise, and to alert the observers - perhaps by triggering a flashing light or an audible alarm - for them to concentrate on dealing with the incident. The trigger could also be used to automatically start up a time-lapse video recording or perhaps increase the frame rate of an ongoing recording. An overview of a possible system is shown in Fig. 1.

The situation is rather different when collecting data to derive information of interest to architects and town-planners. The problem is not time-critical and the processing can be done off-line. The objective is to derive mathematical models of the behaviour of crowds which can form a basis for useful predictions about crowds. For high density crowds, a major difficulty is validation of the models. It is difficult to estimate the actual density or velocity of real crowds. Human observers can be, and are, used to try to determine this data by watching recorded video sequences, but this is a time-consuming and difficult process. There is thus a considerable benefit from being able to develop methods of automatically collecting this data by the use of image processing techniques applied to the video sequences. Provided that such methods can be adequately validated for accuracy by comparison with manual observations or other alternatives, they can be used as a basis for deriving the mathematical models. Our objective for the models is that they should not involve actual counting of individuals or tracking of the movements of individuals but should be based on a collective description of crowds (e.g. analogous to the ideal-gas theory which ignores individual molecules). Therefore, our work has deliberately not attempted to segregate and identify individuals in video images of crowds.

If specially designed and located cameras and other electronic equipment could be installed in all the experimental sites, much better data could be collected. However, this would involve far greater cost than using the existing (and increasing) installed base of closed-circuit TV cameras provided primarily for security reasons. Although these cameras are often not located in the best positions for data collection, the authors believe that any solution must take this into consideration, if it going to be widely applicable.

## Image Processing

Human observers of crowds, particularly those experienced in the management of crowds in public places, can detect many crowd features, in some cases quite easily. Normally they can distinguish between a moving and a stationary crowd and estimate the majority direction and speed of movement of a large crowd, without needing to visually identify or count the separate individuals forming the crowd. They could also easily estimate in a qualitative way the crowd density. The 'mood' of a crowd would also be apparent to a human observer (especially if the vision was augmented by sound).

To require image processing and computer vision techniques to match such capabilities of human observers is at present unrealistic. However, study of the methods used by human observers may help in the choice of image processing algorithms likely to be useful in automatic assessment of crowd behaviour.

### *Experimental set-up*

The development environment consists of equipment for on-site recording (VCR and a camera) and an image processor hosted by a workstation for real-time implementation and algorithmic development (Fig. 2). The image processor contains a network of transputers and a dedicated frame grabber/processor capable of handling up to four monochrome video signals. Initial development was carried out using peak-time recordings made at Liverpool St. railway station (a major commuter station in London, UK) using two camera positions typical of conventional surveillance CCTV equipment. The video images are digitised at a resolution of 512x512 pixels and 256 grey levels. Fig. 3 shows a typical digitised image with about 17 people within an "area of interest" (AOI), shown by a white rectangle.

### *Detection of Stationary Crowds*

It is well-established that crowd congestion which is reaching a danger-level can be spotted by observers noting that the up-and-down oscillatory head movements of individuals walking in a freely-flowing crowd stop when the crowd is too dense for free movement. While working with the authors, Hentschel [2] investigated frequency-domain techniques to identify these up-and-down movements to discriminate between stationary and non-stationary flow. A possible alternative is to compute the 2-dimensional Discrete Fourier Transform (DFT) for each image in a time sequence followed by a measurement of temporal changes in the resulting magnitude and/or phase spectra (frequency domain). This approach has two main disadvantages. First, the DFT for a single image is related to local changes of intensity and not to temporal (interframe) properties. Secondly, it involves a high computational and memory cost. A more effective method is to isolate motion properties in the image sequence through a data-reducing coding mechanism, such as the Discrete Cosine Transform (DCT) whose form for a one-dimensional “image” $f(x,t)$ of N elements is given by:

$$g(t) = \sum_{x=0}^{N-1} f(x,t)\cos(2\pi kx)\,, \qquad (1)$$

where $k$ is a constant derived from the maximum signal frequency (Nyquist criterion) and the maximum expected motion to be observed. This transform associates sinusoids with the time-varying parts of an image (faster moving objects are associated with high frequency sinusoids and non-moving objects are associated with constant levels), which can be detected by applying the DFT on $g(t)$. For one-dimensional “images” of N pixels, the DCT calculates a *single* value for each image in a time-sequence (a data reduction of N to 1), therefore reducing significantly the computational cost of the DFT. As we are only concerned here with detecting movement in the vertical direction, the DCT of a two-dimensional image sequence is given by:

$$g(t) = \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} \mathbf{I}(x, y, t) \cos(2\pi k y) \tag{2}$$

where $\mathbf{I}(x, y, t)$ represents each image (of size NxN) in the sequence. Hentschel [2] developed and tested a number of variants of this scheme, finally proposing the “Linear Area Transform” (LAT) algorithm that takes an image sequence $\mathbf{I}(x, y, 0)$, $\mathbf{I}(x, y, \Delta T)$, $\mathbf{I}(x, y, 2\Delta T)$, ... $\mathbf{I}(x, y, m\Delta T)$ and first computes and *interframe* sequence $\mathbf{D}(x, y, 0)$, $\mathbf{D}(x, y, \Delta T)$, $\mathbf{D}(x, y, 2\Delta T)$, ..., $\mathbf{D}(x, y, (m-1)\Delta T)$, by pixel-to-pixel subtraction between adjacent frames i.e. $\mathbf{D}(x, y, 0) = \left|\mathbf{I}(x, y, \Delta T) - \mathbf{I}(x, y, 0)\right|$, etc. Non-zero pixels in the interframe sequence thus corresponds to areas of movement in the images. The LAT then computes the scalar time sequence:

$$g(t) = \sum_{y=0}^{N-1} y \sum_{x=0}^{N-1} \mathbf{D}\left(x, y, t\right) \tag{3}$$

where NxN is the image size. In other words, the total amount of motion in each image, weighted by vertical position, is accumulated in $g(t)$. To improve peak detection, the d.c. component of $g(t)$ is removed by computing its derivative (difference) to which the DFT is applied. The frequencies of head oscillations correspond to peaks in the resulting frequency spectrum. Fig. 4 shows a typical result for a sequence with 32 images, where the peak at about 2Hz agrees with manual observations.

### Estimation of Crowd Density

Human observers could be expected to estimate crowd density by counting individuals, making use of their ability to rapidly identify the separate individuals using higher-order knowledge about the shape and characteristics of humans. For example the heads of people in a crowd are easy for a human observer to locate, and so counting heads would be a natural approach.

To expect an image processing and computer vision system to identify each individual in a crowd as a preliminary step to counting and tracking is unrealistic at present. Although many sophisticated systems have been devised for military applications (e.g. identifying and tracking multiple targets and decoys) and for some civil applications (vehicle tracking and identification on motorways, pig identification, etc.) there is not yet any realistic possibility of using such methods economically for crowd-behaviour studies. In any case, the main theme of our research is to find general models of crowd behaviour which do not rely upon detecting the behaviour of individuals.

### *Background Removal*

It is clear that a human observer can easily distinguish a very dense crowd from the background (surrounding buildings, road surfaces, and so on) and would be likely to use the ratio of 'crowd area' to 'background area' as a rough estimate for the crowd density. This idea could be applied quantitatively to computer-based density estimation if the image-pixels corresponding to the crowd could be separated from those of the background. This can be done quite effectively using a 'reference image' of the scene, obtained with no crowd present (Fig. 5) for subtraction from the image under analysis. A typical result is shown in Fig. 6. Care has to be taken that lighting conditions are similar, and that there are not other movable objects in the scene (such as vehicles, temporary bill boards or other signs, etc.) which the computer would not distinguish from people.

Slowly-changing lighting conditions can be compensated for by adaptive techniques (for example, by using the mean intensity over the whole image as a signal to control amplifier-gain in the image processor).

Objects other than people could be dealt with if it is considered that the system is really intended to measure floor-occupancy measured in terms of person-equivalents. The unlikely appearance of an elephant in a railway station concourse could then be allowed for by regarding it as a (large) number of person-equivalents in a crowd-density assessment. The requirement is then for human and automatic assessment to give a similar estimate. Fig. 7 shows the relationship between the number of pixels after background removal and manually counted pedestrians for a sequence of more than 150 images (obtained at a frame interval of 10 seconds so that different pedestrians appear in each image). The difference between the measured data and the best fit has a standard deviation of 1.1 pedestrians, equivalent to a 2$\sigma$ relative error of 15% for 15 pedestrians, the level at which crowding starts for the observed area, according to the classification proposed by Polus *et al*. [3]

### *Edge Detection*

An alternative idea is to measure the total perimeter of all the regions occupied by people. For low-density crowds, this can be expected to give a measure of density, although errors are inevitable as numbers increase because of occlusion and overlapping of individuals. Edge-detection is a standard low-level image processing function which can be used to derive outlines of individuals and groups in a video image. The process can be refined further by thinning the edge images to minimise the effects of varying edge thickness. Fig. 8 shows a typical "thinned edges" image.

Fig. 9 shows the relationship between the number of pixels after edge detection/thinning and manually counted pedestrians for a sequence of more than 150 images. The difference between the measured data and a straight line approximation obtained by a least squares fit has a standard deviation of 1.7 pedestrians. This is equivalent to a 2σ relative error of 23% for 15 pedestrians, clearly less accurate than the previous method.

### *Optimal density estimate*

Each of the measurement techniques proposed here (background removal and edge detection) can be approximated by the linear relationship

$$z = mx + b \qquad y = (z - b) = mx \tag{4}$$

where *z* is the number of pixels after segmentation (e.g. number of non-background pixels or number of thinned edge pixels), *x* is the number people and *m* and *b* are coefficients obtained from the experimental data by linear regression. Therefore, it is possible to combine these two measurements into an “optimal” estimation of crowd density through a linear Kalman filter [4]. A simple dynamic model is used [5] that assumes a constant number of pedestrians where the actual variation of pedestrian density from image to image is modelled as zero-mean “process” noise, as follows:

$$\begin{aligned} &x_{k+1} = x_k + v \\ &\left| \begin{matrix} y_e \\ y_b \end{matrix} \right|_k = \begin{bmatrix} z_e - b_e \\ z_b - b_b \end{bmatrix} = \begin{bmatrix} m_e \\ m_b \end{bmatrix} x_k + \begin{bmatrix} w_e \\ w_b \end{bmatrix} \end{aligned} \tag{5}$$

where *x* is the number of people, *k* is sample (image) number, $v$ is "process" noise (obtained from the variation of *x* from image to image), *z* is the number of *crowd* pixels, *m* and *b* are the linear-fit coefficients and *w* represents the noise characteristics of each measurement technique (subscripts *e* and *b* indicate thinned edges and background removal respectively). When operating at a frame interval of 10 seconds, application of the filter results in a mean relative error (compared with manual counts) of less than 8%. This is illustrated in Fig. 10 (for clarity, the manual counts have been shifted vertically by 10 units). In the existing system it is possible to calculate densities at a rate of about 2 frames per second. A typical update rate for operators is 10 seconds, giving 20 measurements for Kalman filtering and thus an accuracy figure better than the one quoted above should be expected.

### *Geometric Distortion*

Both of these, and other, methods suffer from a near-far effect, since people near to the camera will occupy more area than people of the same size far from the camera. The problem is exacerbated if standard closed-circuit security camera installations are used for data-capture, because such cameras are commonly mounted at low angles either to obtain a longer field of view for human monitoring or because they are installed in locations with low ceilings (e.g. tunnels or platforms in subway areas). Compensation for this near-far effect can be partially achieved by using geometrical distortion of the image (Fig. 11). Such distortion is commonly used for special video effects in the television entertainment and advertising business, and special hardware is available for rapid processing. The technique also has applications for video data-compression, and so the theory and practice is well developed.

## Estimation of Crowd Motion

Human observers have a highly developed capability for visual tracking of moving objects in a complex scene, which is not easily matched by computers. In semi-confined spaces individuals are free to move in various directions. Moreover, at any one instant, different parts of a single individual (e.g. head, limbs) move in different ways.

Recognising movement in a sequence of video images, without identifying objects or "understanding" the scene, requires a technique which tracks the displacement (and hence the velocity) of regions of similar brightness-patterns from one image to the next. Of course, such velocity estimation cannot distinguish between individuals or separate the movement of humans from other moving objects. However, crowd analysis is usually more concerned with group behaviour such as preferential motion direction and magnitude. For instance, a typical useful measure is the distribution of the proportion of people moving in a discrete set of preferential directions (e.g. architects normally use a "wind rose" of eight directions: North, North East, etc.). The approach proposed here is therefore to measure motion features at pixel or pixel-neighbourhood level which are then aggregated to obtain motion properties for larger regions in an image. The aggregated results can then be used to establish overall preferential crowd velocities (direction and magnitude).

### *Conventional Optical Flow Computation*

Motion can be measured by computing the “optical flow” (defined as the change of image brightness from one image to the next expressed in terms of a vector field, and resulting from the projection of the true 3-dimensional velocity field on the image plane [6). If an image is brightness at time $t$ is represented by $\mathbf{I}(x, y, t)$ , then:

$$\frac{d\mathbf{I}}{dt} = \frac{\partial \mathbf{I}}{\partial x}\frac{dx}{dt} + \frac{\partial \mathbf{I}}{\partial y}\frac{dy}{dt} + \frac{\partial \mathbf{I}}{\partial t} = \frac{\partial \mathbf{I}}{\partial x}u + \frac{\partial \mathbf{I}}{\partial y}v + \frac{\partial \mathbf{I}}{\partial t} \qquad (6)$$

The objective is to calculate the motion vectors ($u$, $v$) for all points in the image. A well-known method to solve the above equation is based on the so-called Optical Flow Constraint (OFC), proposed by Horn and Schunck [7] which assumes that the brightness is constant with respect to time (i.e. $d\mathbf{I}/dt = 0$). This assumption is difficult to satisfy in the presence of sudden illumination changes or occlusion (i.e. discontinuities). However, the method can be shown to be useful for separating moving from stationary crowds (a useful indicator of congestion or potential danger). Improved results and computation times are obtained if the background is removed from the two successive images to pre-select pixels of interest (i.e.

pedestrian pixels). After calculating the optical flow for these pixels, small disjoint neighbourhoods are used to average the vector field. This reduces noise and provides a better visualisation of the field. Fig. 12 shows a typical result for a case of a standing queue (right hand side) and moving pedestrians (left hand side). The average optical flow vectors have been superimposed on the original image (the small white square indicates the origin of each vector).

*Block-matching motion detection*

It is well known that Horn's optical flow algorithm is limited to small pixel displacements. An alternative way to detect motion is by identifying pairs of pixel neighbourhoods, in the two successive images, that have a similar grey level distribution. Pixels in the first image can, as before, be preselected to reduce the amount of data to process. This can be done using background removal (pixels in the background are not expected to move) or by computing the difference between the two images (a direct indication of motion).

The objective is then to compute a velocity field for each preselected pixel (x,y) in the first image by defining a small neighbourhood (or *pixel block*, typically 10x10) centred at position (x,y). A *search* block, also centred at (x,y), is defined in the second image. The size of the search block is determined by the maximum displacement expected in the given frame interval. Then, all possible pixel blocks in the search area are compared with that in the first image using a similarity function (e.g. sum of the pixel-to-pixel absolute difference). Under ideal conditions of no changes in illumination and object shape, a *matching* block would be found where the similarity function is zero. Under more realistic conditions, a match is defined as the block that produces the minimum similarity value. A threshold is applied to account for cases of drastic changes in shape or illumination or that of objects leaving the scene.

As an example, Fig. 13 shows a pixel block in a first image and a search block in the second image (sizes have been exaggerated for illustration purposes). The two images are separated by an interval of 0.12 seconds.

Irregular motion, movements of arms, legs, and clothing and localised variations in brightness all cause errors in the computed motion vectors compared to the actual overall motion of the individuals in the crowd. To compensate for such effects (which can be effectively regarded as zero-mean noise added to the motion vector estimates), the computed vectors are aggregated over small disjoint neighbourhoods of 10x10 pixels throughout the image. A typical resulting vector field superimposed on the first image is shown in Fig. 14. Reference to the original images confirms a correlation between these measurements and manual observations. Pixel pre-selection by background removal, and to a lesser extent interframe difference, results in a reduced number of mismatches compared to pre-selection based on thinned edges [8,9].

The motion vectors calculated by block matching and averaging may be used to devise a polar plot (showing velocity magnitude *s* and direction *θ*) for a moving crowd. A typical plot is shown in Fig. 15, for pixel preselection based on background removal.

Improved human and machine interpretation of these polar diagrams is achieved when the velocity vectors are aggregated within discrete direction "bins" (in a polar histogram). Fig. 16 shows such a plot for the data in Fig. 15 for a direction bin-size $\Delta\theta = 1°$, where the dominant south-east motion tendency can be clearly seen.

The block matching technique described here involves substantial computation. However, the commercial demands for video-data compression are such that specially-designed VLSI components are already being developed for fast implementation of these algorithms (e.g. SGS-Thomson's STI-3220 Motion Estimation Processor). It is therefore feasible to use these techniques for real-time crowd velocity estimation.

## Conclusions

This article has shown that it is possible to use well-established image processing techniques for monitoring and collecting data on crowd behaviour. A key factor in the solutions described is the use of global or semi-global pixel intensity values to infer crowd behaviour avoiding recognition and tracking of individual pedestrians. The methods discussed are amenable to real-time implementation.

## Acknowledgements

The authors are grateful to the EPSRC for their financial support, to their colleagues Maria Alicia Vicencio-Silva, Richard Allsop and Alan Penn (University College London) and to Railtrack plc and London Underground for granting access to their sites and their expertise.

## References

1. Hillier B., Penn A., Hanson H., Grajewski T. and Xu. J, 1993, “Natural movement: or, configuration and attraction in urban pedestrian movement”, Environment and Planning B, 20, 29-66
2. Hentschel T., 1993, "Image Processing Techniques for the Estimation of Features of Crowd Behaviour in Urban Environments", MSc. Dissertation, King's College London, UK
3. Polus A., Schofer J.L. and Ushpiz A., 1983, “Pedestrian Flow and Level of Service”. J. Transportation Engineering, 109, 46-56
4. Brown R.G. and Hwang P.Y.C., 1992, "Introduction to Random Signal Analysis and Kalman Filtering", Wiley, 2nd ed.

5. Velastin S.A., J.H. Yin, A.C. Davies, M.A. Vicencio-Silva, R.E. Allsop R.E., and A. Penn, 1993, "Analysis of Crowd Movement and Densities in Built-up Environments using Image Processing", IEE Coll. on Image processing for Transport Applications, 9 Dec., London, UK. Digest No. 1993/236, 8/1-8/6
6. Ben-Tzvi D., Del Bimbo F. and Nesi P., 1993, “Optical flow estimation by using the Combinatorial Hough Transform”, Procs. of the 7th International Conference on Image Analysis and Processing, Bari, Italy, 20-22 Sept. 1993, 531-538
7. Horn B.K.P. and Schunck B.G., 1981, “Determining optical flow”. Artificial Intelligence, 17, 185-203
8. Velastin S.A., Yin J.H, Davies A.C., Vicencio-Silva M.A., Allsop R.E. and Penn A., 1994: "Automated Measurement of Crowd Density and Motion using Image Processing", 7th IEE International Conference on Road Traffic Monitoring and Control, 26-28 April 1994, London, UK, 127-132
9. Velastin S.A., Yin J.H., Vicencio-Silva M.A., Davies A.C., Allsop R.E. and Penn A., 1994: "Image Processing Techniques for On-line Analysis of Crowds in Public Transport Areas", IFAC Symposium on Transportation Systems: Theory and Application of Advanced Technology, 24-26 August 1994, Tianjin, China

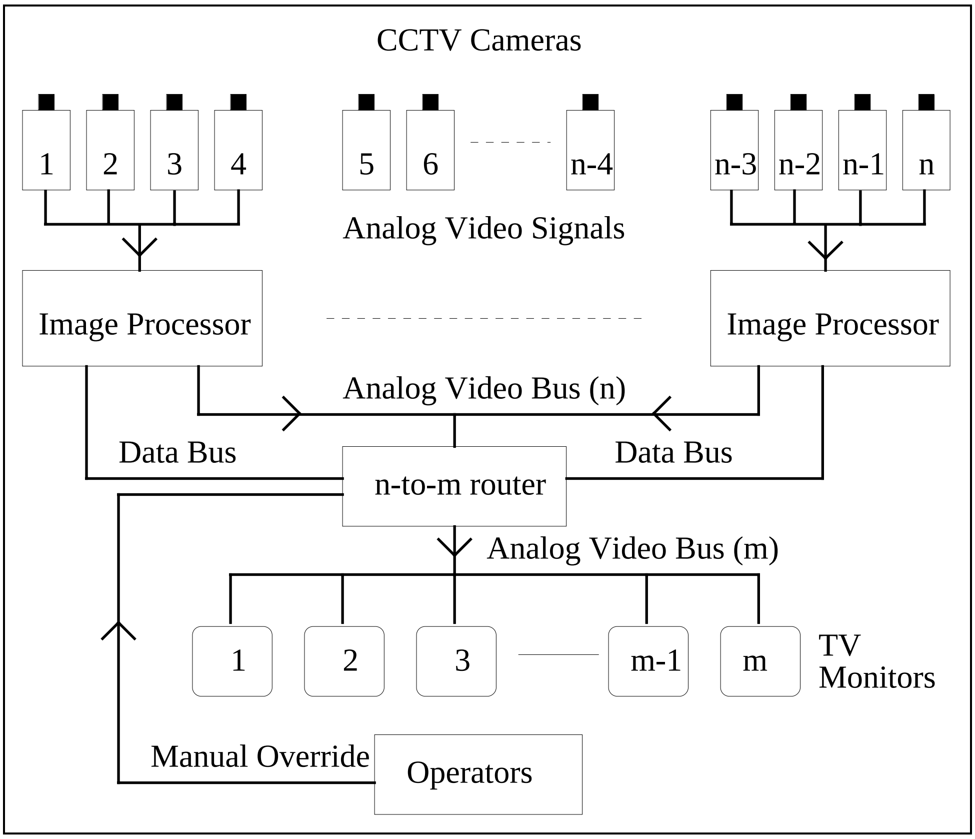


Fig. 1: Overview of a possible system: Images from all cameras (N) are processed by computers to select those of most interest to the operators (M monitors) e.g. for alerting them of a potentially dangerous situation.

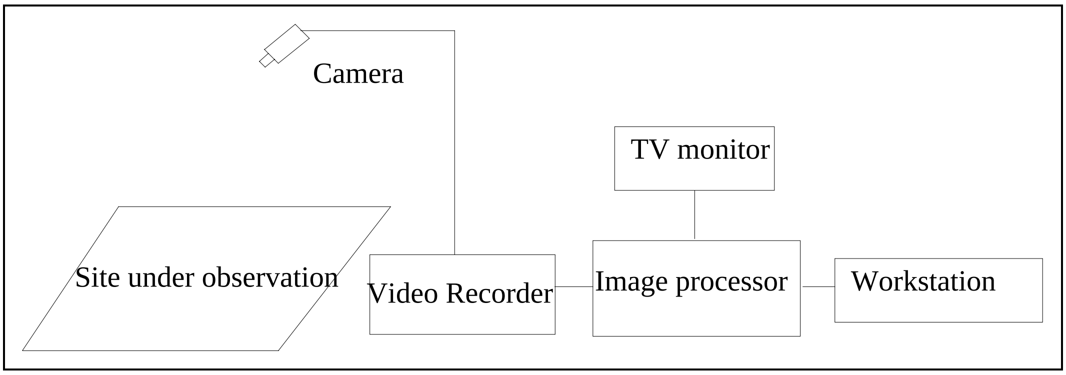


Fig. 2: Development Environment: A standard domestic camera (camcorder) is used to record scenes in a site of interest. The recordings are digitised and processed by a transputer-based image processor under the control of a standard workstation (PC or Sun).

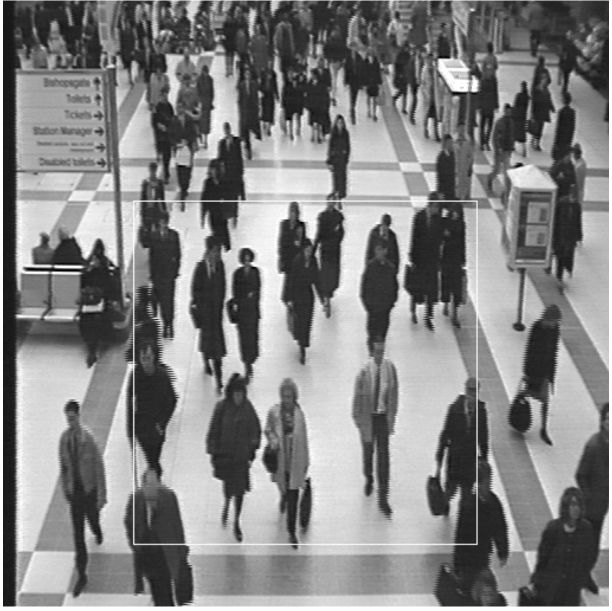
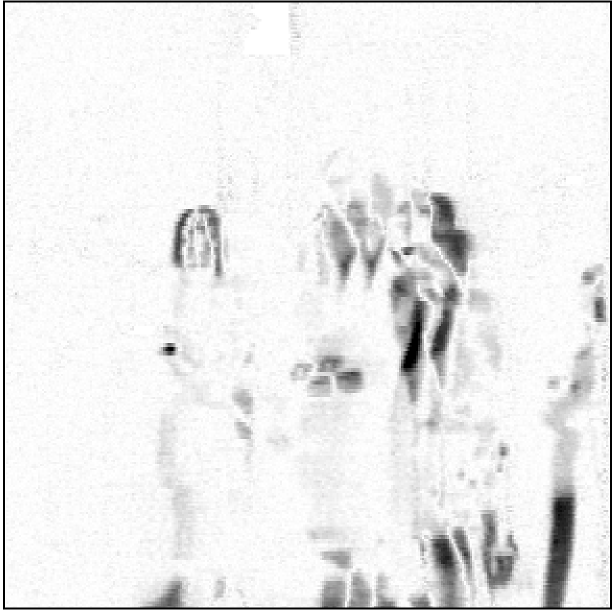
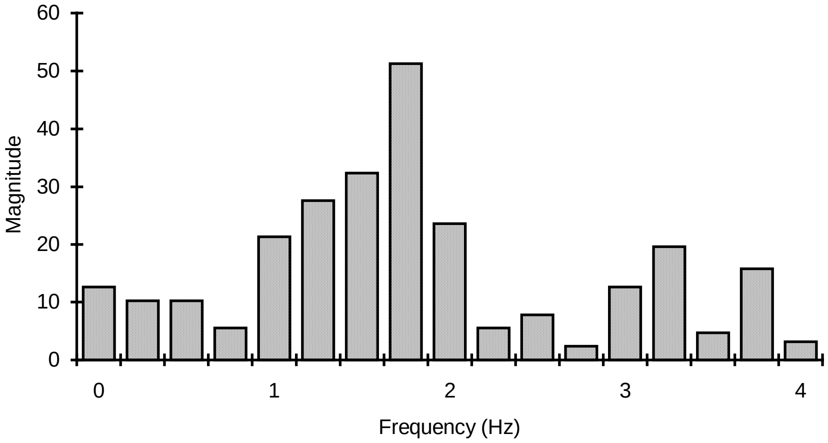


Fig. 3: Typical digitised image (Liverpool St. Railway Station, courtesy of Railtrack East Anglia), 512 x 512 pixels, 256 gray levels.

(a)



(b)



(c)

Fig. 4: a) Typical digitised image (Liverpool St. Underground Station, courtesy of London Underground), b) Difference between two successive images. The result indicates the areas in the image where motion occurs,

c) Frequency spectrum of $\frac{dg(t)}{dt}$ for the image in (b), the peak at about 2Hz corresponds to vertical body motion, indicating that the crowd is moving.

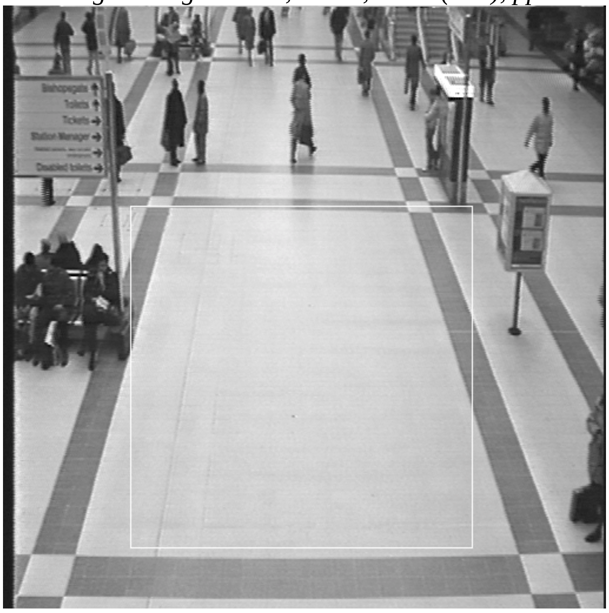


Fig. 5: A "background-only" image for the site in Fig. 3. This is used to isolate pedestrians from images by removing the surrounding background

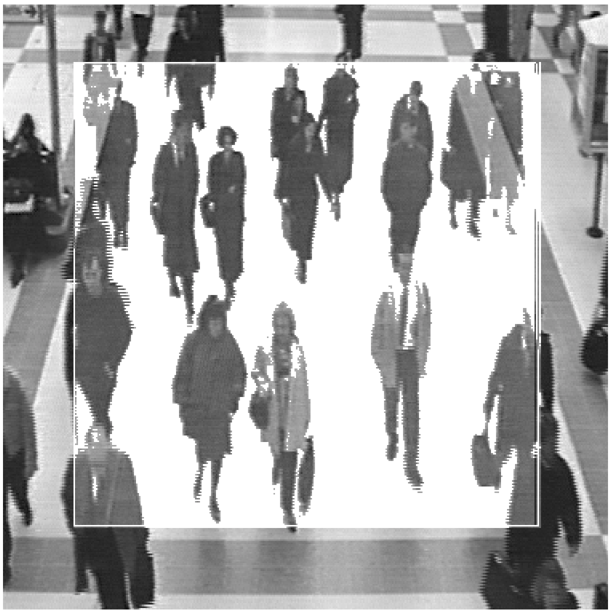


Fig. 6: An image where the background has been removed. The number of remaining "picture elements" has been found to be correlated to the number of people in the image (see Fig. 7)

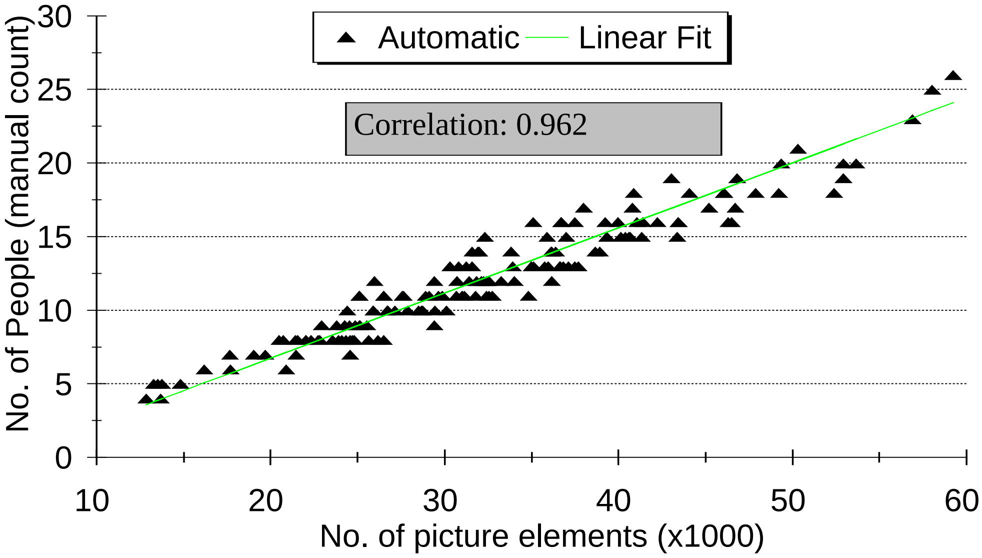


Fig. 7: Relationship between number of people in an image (counted manually) and the number of picture elements after background removal. Each represents data from one image of a sequence of 150 images taken at 10 sec. intervals.

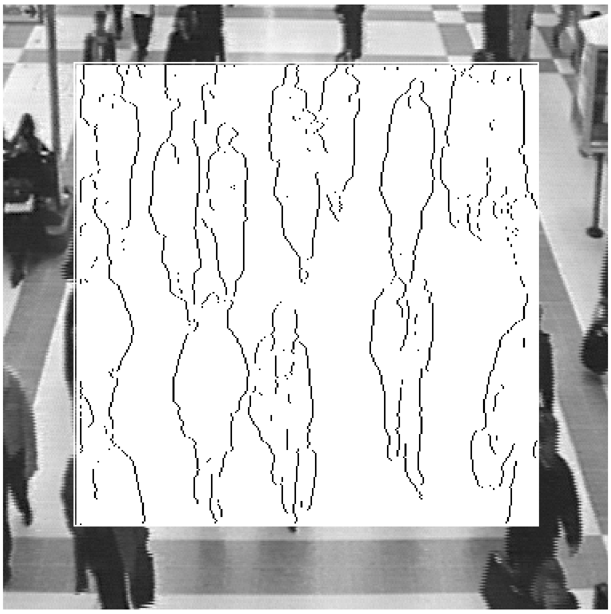


Fig. 8: An image where pedestrian “edges” have been extracted and thinned. The number of remaining “picture elements” has been found to be correlated to the number of people in the image (see Fig. 9)

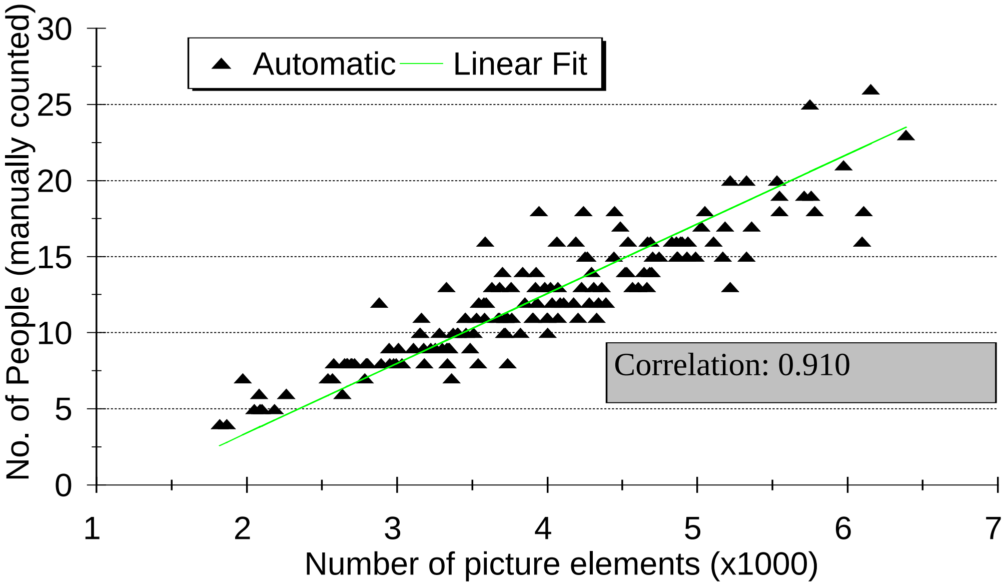


Fig. 9: Relationship between number of people in an image (counted manually) and the number of picture elements after extracting and thinning pedestrian edges. Each represents data from one image of a sequence of 150 images taken at 10 sec. intervals

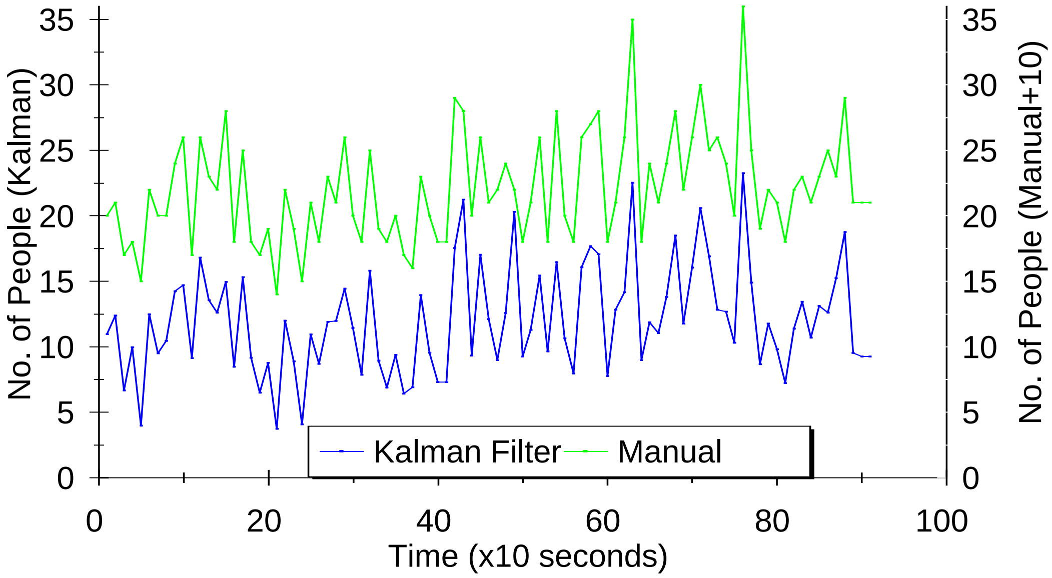


Fig. 10: The "background removal" and "pedestrian edges" techniques are combined into a single measurement using optimal Kalman filtering. The diagram shows automatic (bottom trace) and manual pedestrian counts (top trace) for an image sequence (the manual counts have been shifted by ten vertical units for clarity). The automatic counts follow manual counts closely (within a mean error of 8%).

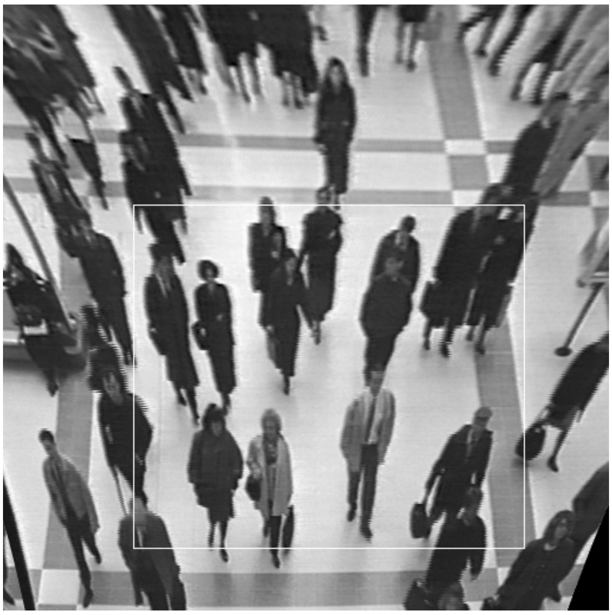


Fig. 11: Geometric distortion of image in Fig. 3. This produces the optical illusion of a vertical camera and simplifies camera calibration for different camera positions.

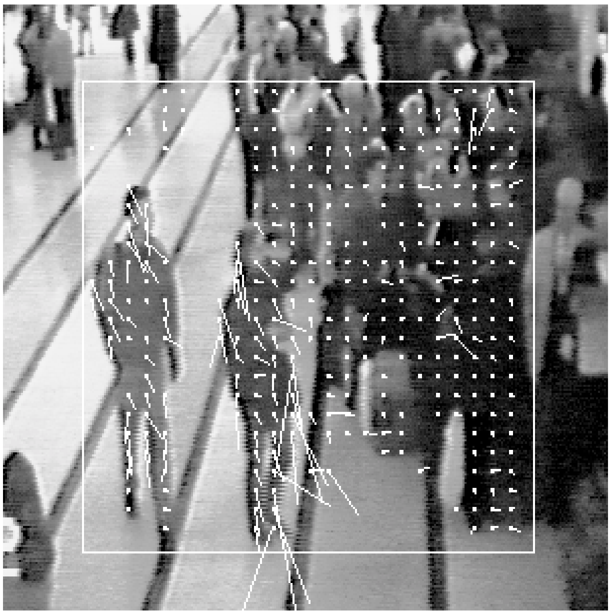


Fig. 12: Areas of movement are detected by “optical flow” calculation. The white “needles” superimposed on the image represent magnitude and direction of movement. The example shows that it is possible to distinguish between stationary (the queue on the right) and moving people (King’s Cross railway station, courtesy of Railtrack plc)

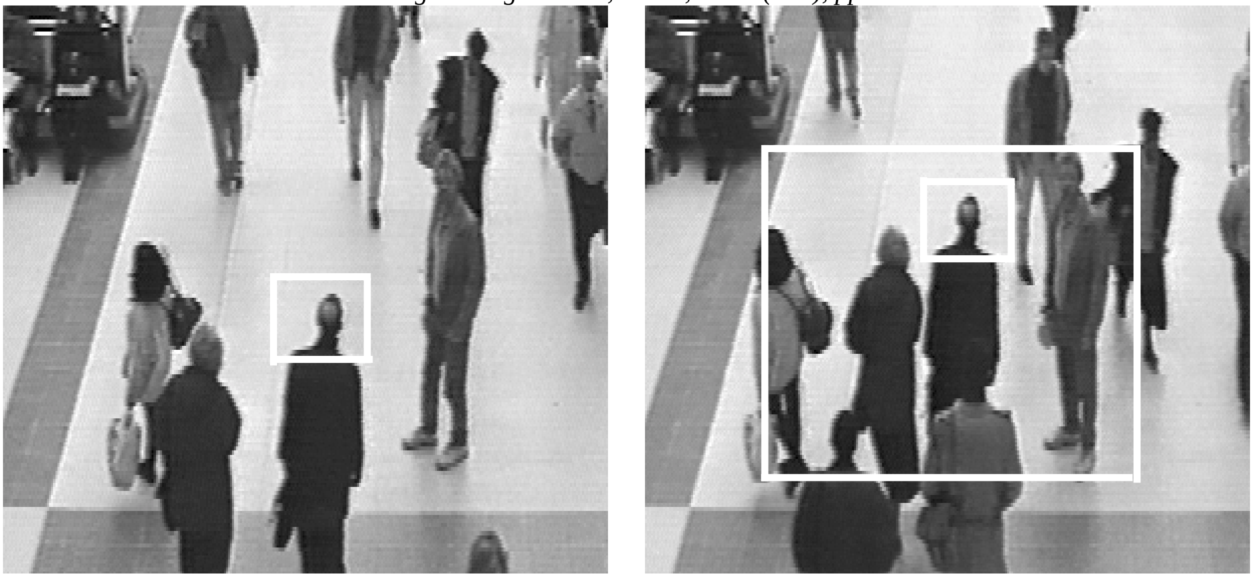


Fig. 13: Motion calculation by “block matching”. A small block is selected as shown on the left image. A “search area” is selected in the next image (on the right) within which a similar block is found. The displacement of the small block between the two images defines a motion vector (see Fig. 14)

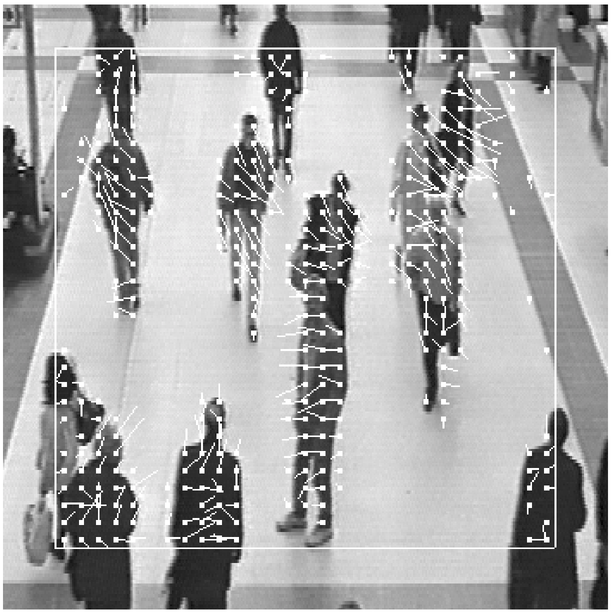


Fig. 14: A typical result after calculating motion by “block matching”. The white “needles” superimposed on the image represent average magnitude and direction of motion.

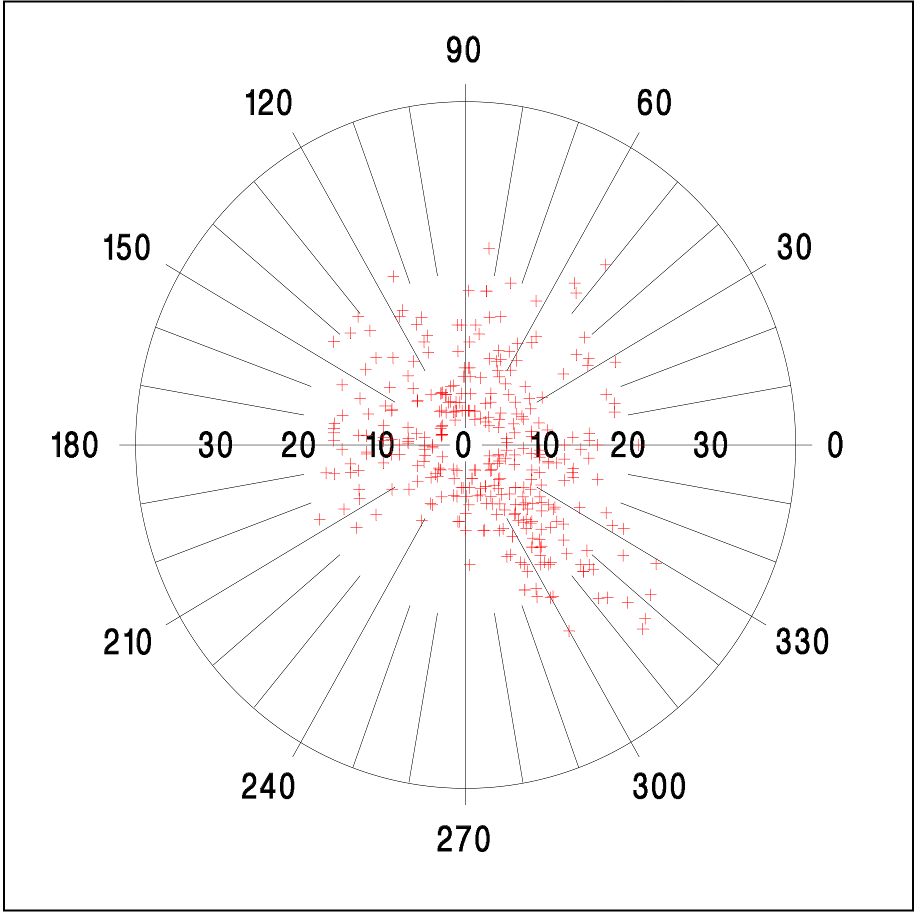


Fig. 15: Motion vectors represented in a "radar" plot (magnitude: radius, direction: angle) for the block matching technique, using background removal to pre-select pixels of interest.

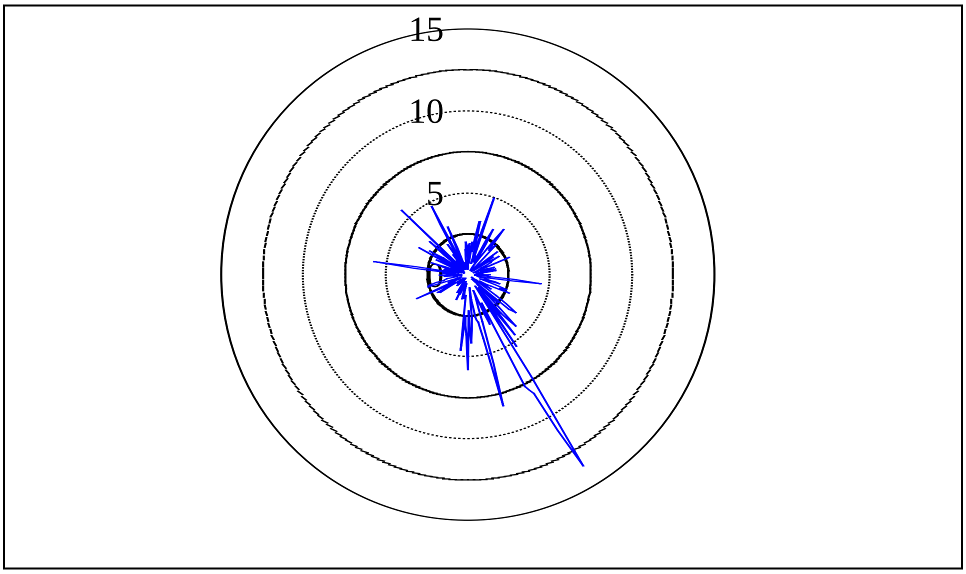


Fig. 16: Motion vectors in an image (background removal). Magnitudes are added in direction intervals of 1 degree. In the example, the main south-east motion tendency can be seen.