第 29 卷 第 3 期
2003 年 5 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 29, No. 3
May, 2003

Learning Surveillance Tracking Models for the Self-Calibrated Ground Plane

J. Renno P. Remagnino G. A. Jones
(*Digital Imaging Research Centre, Kingston University, Kingston, U. K.*)
(E-mail: {j. renno, p. remagnino, g. jones}@kingston. ac. uk)

Abstract We propose a novel method for combining information streamed by a multi-sensor system for visual surveillance. Information fusion occurs in two phases during which all cameras are calibrated with respect to independent global Cartesian reference frames (set on the ground plane) and then all frames are registered into a single coordinate system. The development of automatic calibration and registering of visual data is crucial in visual surveillance applications because it makes easier to install the monitoring infrastructure and, consequently, to develop more accessible Visual Surveillance tools for the public domain. Machine learning techniques are believed to offer the best mathematical tools to handle the uncertainty and incomplete nature of surveillance data.

Key words Visual surveillance, machine learning, data fusion, camera calibration

1 Introduction

By far the most common approach to tracking in typical surveillance imagery uses pixel differencing and blob analysis. Typically motion detection extracts moving regions from static scenes[1]. Trajectory tracking is employed to establish the temporal history of individual objects. An iterative estimator (e. g. Kalman or α-β) is employed to update a first or second order visual trajectory model. Temporal correspondence (or data association) is achieved essentially using simple Newtonian physics either locally for each object, or globally by considering all possible object-observation pairings[2]. Additionally an appearance model matching may be employed to improve tracking accuracy by comparing width and height, shape or colour[3, 4]. While surprisingly successful, maintaining temporal correspondence is a significant problem particularly through occlusion and fragmentation where the shape, dimensions and colour signature of the merged or fragmented observations do not correlate well with the actual object, or where the trajectory model does not correspond to actual object trajectory.

Two related problems are addressed in this paper. First the problem of frequent fragmentation and merging of moving regions caused by occlusion and low contrast processes. These unexpected regions usually introduce considerable noise into the data association phase of the tracker and, more subtlety, into the updating of the trajectory and appearance models which is then propagated into the subsequent frames. Typical solutions are complex and ad hoc split and merge procedures applied to observation and appearance model primitives[5~7].

The second problem relates to the choice of motion model. Linear pixel-based motion models applied to trajectory and appearance models are too constrained to adequately model the evolution of objects - particularly vehicles. The result is frequent loss of correspondence as objects manoeuvre in the scene. On the other hand, more appropriate quadratic models are easily mislead by observation noise. The difficulty lies in the problem of establishing global pixel-based noise models which are appropriate to both distant visually-slow objects at the top of the image, and closer objects with larger visual velocities at the bottom.

Received November 30, 2002; in revised form April 7, 2003
收稿日期 2002-11-30; 收修改稿日期 2003-04-07

In this work, we introduce three mechanisms to address these problems which rely on knowledge about the ground plane. First we develop highly discriminatory bounding-box appearance models of scene objects which indirectly use the depth of the object to model its projected width and height. Since, the spatial extent of an object is now a function of image position, the tracker will be more robust when presented with the distorted observations which arise from fragmentation or occlusion processes. Second, the observations are transformed onto the ground plane co-ordinate system within which a quadratic rather than linear motion model is defined. Global real-valued noise models can be generated for observation and dynamic noise models. Finally, rather than relying on a labour-intensive calibration procedures to recover the image to ground-plane homography[8], the system relies on a simple auto-calibration procedure to learn the relationship between image and world by simply watching events within the monitored scene. Having calibrated each camera to its local ground plane, Section 4 demonstrates how these ground planes may be registered. Again a learning procedure is pursued in which the projected trajectory positions in corresponding frames and their instantaneous velocity estimates are combined to create estimates of the rotation and translation. A clustering algorithm is used to locate the most likely transform between each pair of camera ground planes.

2 Auto-calibration of the ground plane

In this section a simple yet highly effective method of learning the image to ground plane homography of the camera is presented which exploits the simple but reasonably accurate assumption that in typical surveillance installations, the projected 2D image height of an object varies linearly with its vertical position in the image - increasing down the image from zero at the horizon. This height model is derived from the optical geometry of a typical visual surveillance installation. In addition, such an assumption enables the use of simple but highly discriminatory models of the appearance of scene objects which indirectly use the depth of the object to model its projected height. In this auto-calibration scenario, the ground plane co-ordinate system (GPCS) is defined as follows:

The Y -axis $\hat{\boldsymbol{Y}}$ of the GPCS is defined as the projection of the optical axis along the ground plane. The Z -axis $\hat{\boldsymbol{Z}}$ is defined as the ground plane normal. The position of the camera focal point in the GPCS is 'above' the GPCS origin at the point $(0, 0, L)$.

2.1 Ground plane projection

The image plane is situated at distance f (focal length of the optical system for the camera) perpendicular to the optical axis $\hat{z}$. In this configuration a point $\boldsymbol{P}$ on the image plane has co-ordinates $\boldsymbol{x}'=(x, y, -f)^{\mathrm{T}}$, where x, y are image plane positions and f, the focal length. The pixel co-ordinate system i, j (representing the row and column position respectively) is related to the image plane co-ordinate system by $x=\alpha_x(j-j_0)$ and $y=\alpha_y(i_0-i)$ where i_0, j_0 are the optical centre of the image and α_x and α_y are the horizontal and vertical inter-pixel widths. Thus

$$\boldsymbol{x}' = (\alpha_x^f(j-j_0), \alpha_y^f(i-i_0), -1)^{\mathrm{T}} f$$

where α_x^f and α_y^f are the horizontal and vertical pixel dimensions normalised by the focal length.

An optical ray containing the focal point of the camera passing through the image plane can be represented in vector form as $\boldsymbol{x}=\mu\boldsymbol{x}'$ where μ projects the point $\boldsymbol{x}'$ arbitarily. Let Q be the point of intersection of the optical ray with the ground plane Π. In order to calculate the position $\boldsymbol{x}$ of the point Q on the ground plane Π in the ground plane co-ordinate system, one must convert the position of an image point $\boldsymbol{x}$ given the transformation $(R, \boldsymbol{t})$ between the image plane and local world co-ordinate systems i. e. $\boldsymbol{X}=\mu R\boldsymbol{x}'+\boldsymbol{t}$. Writing the ground plane equation as $\boldsymbol{n}_\Pi \cdot \boldsymbol{X}=0$, where the ground plane normal $\boldsymbol{n}_\Pi \equiv \hat{\boldsymbol{Z}}$,

then the position $\boldsymbol{X}$ of the point Q is obtained by noting that $\boldsymbol{X} \cdot \hat{\boldsymbol{Z}}=0$.

$$\mu = - t_z / \hat{\boldsymbol{Z}} \cdot R\boldsymbol{x}' \tag{1}$$

The local GPCS is defined with a zero pan angle. Assuming no significant roll angle, then after some algebraic manipulation, the ground plane co-ordinates may be related to the look down angle θ as follows.

$$\frac{X}{L} = \frac{\alpha_x^f (j - j_0)}{\alpha_y^f (i - i_0)\sin\theta - \cos\theta}, \quad \frac{Y}{L} = - \frac{\alpha_y^f (i - i_0)\cos\theta - \sin\theta}{\alpha_y^f (i - i_0)\sin\theta - \cos\theta} \tag{2}$$

Thus to compute the ground plane position of an image point, the following camera parameters i_0, j_0, α_x^f, α_y^f, and θ are needed. In our approach the optical centre i_0, j_0 is computed by an optical flow algorithm which robustly fits a global zoom motion model to a three frame sequence undergoing a small zoom motion.

2.2 Projected object height

If one assumes that the height of a moving object is known (*i.e.* a person) then the point of intersection $\boldsymbol{X}$ with the ground plane can be shifted along the $\hat{\boldsymbol{Z}}$ direction by the height H. Using μ, we can write $\boldsymbol{X}' = \mu R\boldsymbol{x}' + \boldsymbol{t} + H\hat{\boldsymbol{Z}}$. The new image point $\boldsymbol{x}''$ corresponding to the projection of the top of the person can be computed from the inverse transformation $\boldsymbol{R}^{\mathrm{T}}(\boldsymbol{X}' - \boldsymbol{t})$ to yield

$$\lambda \boldsymbol{x}'' = \mu \boldsymbol{x}' + H R^{\mathrm{T}} \hat{\boldsymbol{Z}} \tag{3}$$

where λ is the projection factor from the image plane to the top of the person. Substituting μ from Equation (1) and $t_z = L$ yields

$$\boldsymbol{x}'' = - \frac{1}{\lambda}\left(H\, R^{\mathrm{T}} \hat{\boldsymbol{Z}} - \frac{L}{\hat{\boldsymbol{Z}} \cdot R\boldsymbol{x}'} \boldsymbol{x}' \right) \tag{4}$$

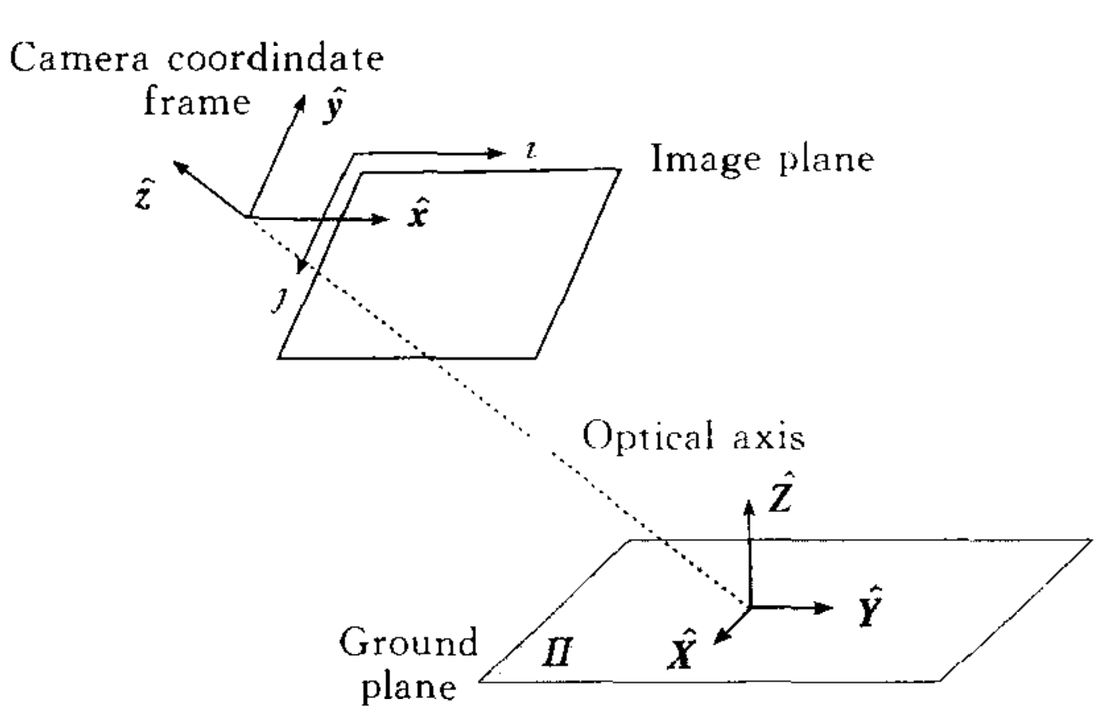
To measure the projected vertical height of an object, we simply define a plane Λ containing the optical centre and the image plane raster line containing the new point $\boldsymbol{x}''$ (see Fig. 1(b)). The normal $\boldsymbol{n}_\Lambda$ of this plane is defined by the cross-product between the projection line $\lambda \boldsymbol{x}''$ and the raster line direction vector $\hat{\boldsymbol{x}}$ as follows

$$\boldsymbol{n}_\Lambda = - \frac{1}{\lambda}\left(H\, R^{\mathrm{T}} \hat{\boldsymbol{Z}} \times \hat{\boldsymbol{x}} - \frac{L}{\hat{\boldsymbol{Z}} \cdot R\boldsymbol{x}'} \boldsymbol{x}' \times \hat{\boldsymbol{x}} \right) \tag{5}$$

The raster line containing the point $\boldsymbol{x}''$ can be thought of as lying at a distance h above the projection of the bottom of the person - see Figure 1(b). Therefore the point vertically above $\boldsymbol{x}'$ can be expressed as $\boldsymbol{x} = \boldsymbol{x}' + h\hat{\boldsymbol{y}}$ and belongs to the plane Λ. Substituting $\boldsymbol{x}' + h\hat{\boldsymbol{y}}$ into the equation of plane Λ, $\boldsymbol{n}_\Lambda \cdot \boldsymbol{x} = 0$ generates

$$h = - \frac{\boldsymbol{n}_\Lambda \cdot \boldsymbol{x}'}{\boldsymbol{n}_\Lambda \cdot \hat{\boldsymbol{y}}} \tag{6}$$

Further simplification can be derived by expanding the numerator and denominator of Equation (6) using Equation (5) as follows



(a) Camera, world and pixel coordinate systems (b) Projected height

Fig. 1 Coordinate systems

$$-\lambda \boldsymbol{n}_{\Delta} \cdot \boldsymbol{x}' = H(R^{\mathrm{T}}\hat{\boldsymbol{Z}} \times \hat{\boldsymbol{x}}) \cdot \boldsymbol{x}' - \frac{L}{\hat{\boldsymbol{Z}} \cdot R\boldsymbol{x}'}(\boldsymbol{x}' \times \hat{\boldsymbol{x}}) \cdot \boldsymbol{x}' \tag{7}$$

$$= H(R^{\mathrm{T}}\hat{\boldsymbol{Z}} \times \hat{\boldsymbol{x}}) \cdot \boldsymbol{x}' \tag{8}$$

since $(\boldsymbol{x}' \times \hat{\boldsymbol{x}}) \cdot \boldsymbol{x}' = 0$, and

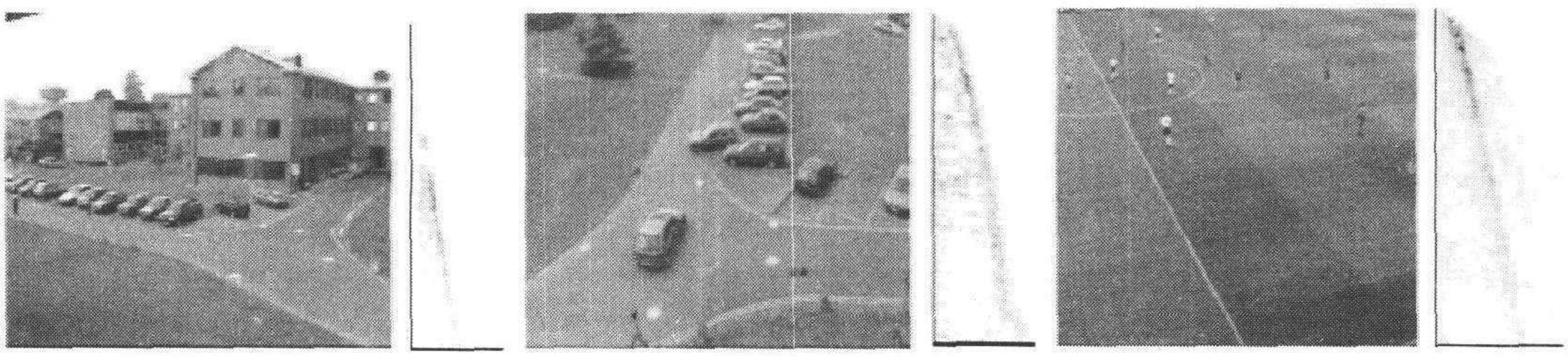
$$-\lambda \boldsymbol{n}_{\Delta} \cdot \hat{\boldsymbol{y}} = H(R^{\mathrm{T}}\hat{\boldsymbol{Z}} \times \hat{\boldsymbol{x}}) \cdot \hat{\boldsymbol{y}} - \frac{L}{\hat{\boldsymbol{Z}} \cdot R\boldsymbol{x}'}(\boldsymbol{x}' \times \hat{\boldsymbol{x}}) \cdot \hat{\boldsymbol{y}}$$

$$= H(R^{\mathrm{T}}\hat{\boldsymbol{Z}} \times \hat{\boldsymbol{x}}) \cdot \hat{\boldsymbol{y}} - \frac{Lf}{\hat{\boldsymbol{Z}} \cdot R\boldsymbol{x}'} \tag{9}$$

where $(\boldsymbol{x}' \times \hat{\boldsymbol{x}}) \cdot \hat{\boldsymbol{y}} = f$. Where there is a zero roll angle, Equations (8) and (9) combine to generate the following expression for image plane height h which depends only on object height H, camera height L and vertical image height y.

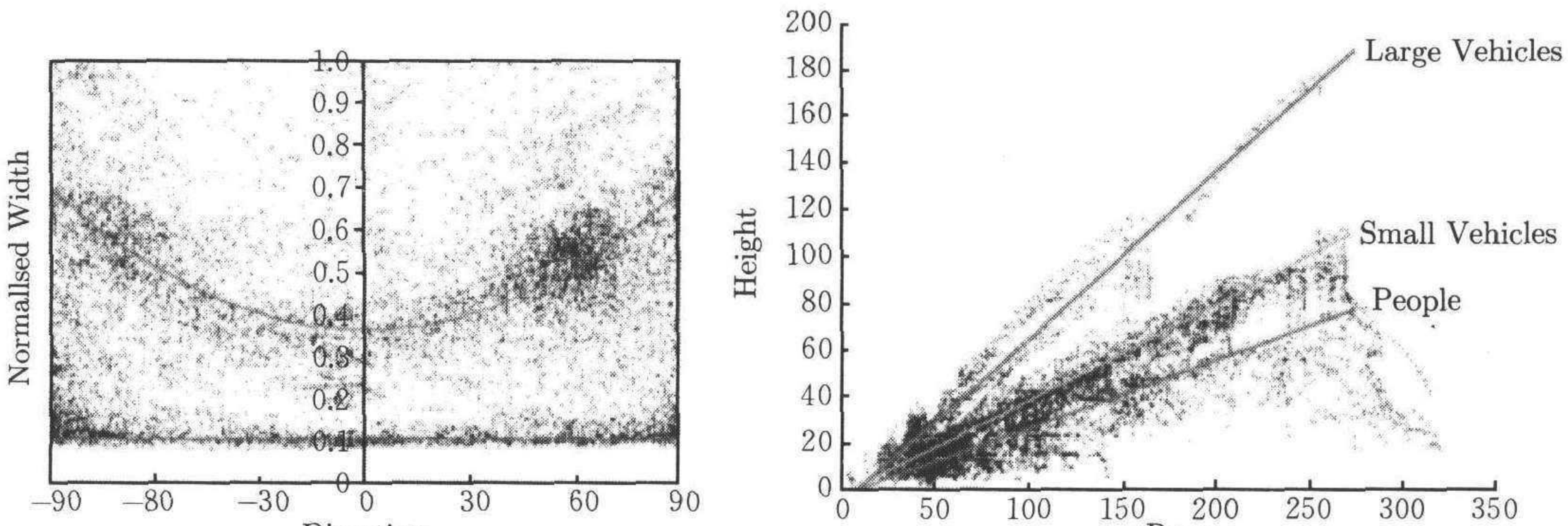
$$h = \frac{(f^2 - y^2)\sin\theta\cos\theta + yf(\cos^2\theta - \sin^2\theta)}{y\sin\theta\cos\theta - (\cos^2\theta - L/H)f} \tag{10}$$

For typical camera installations, h can be shown to effectively vary linearly with vertical image position relative to the position of horizon. The intercept with the vertical position axis (or $h=0$ axis) defines the horizon where objects become infinitely small. Such a linear model may be extracted from moving regions of the monitored scene-see Figures 2[1] and 3. Currently the operator drags a line segment along the ridge structure to define the gradient γ and horizon i_h.



(a) PETS camera 1 (b) PETS camera 2 (c) Football

Fig. 2 PETS cameras 1,2 and football



(a) Normalised width (b) Projected height

Fig. 3 $\theta = 0$ indicates vertical motion, while $\theta = \pm\pi/2$ refers to Horizontal motion. The lower plot illustrates that Person width does not depend on orientation. For Vehicles, the width increases from a minimum at $\theta = 0$ (face on) to a maximum at $\theta = \pm\pi/2$ (side views)

2.3 Ground plane calibration

Since the vertical image height of an object is independent of the horizontal image po-

1) PETS 2001 Datasets visualsurveillance. org & Fulham FC

sition of the projected object, the following derivation may assume, without loss of generality, that the object is located on the vertical axis *i. e.* $x=0$. Two key positions of a projected object may be defined at $i=i_h$ at the horizon, and $i=i_0$ at the optical centre of the image. At the former, the look-down angle θ may directly be related to the horizon parameter i_h extracted from the accumulated training data acquired in the learning stage described in Section 2.2 *i. e.*

$$\cot\theta = \alpha_y^f (i_h - i_0) \tag{11}$$

For the latter case, consider an object of height H standing on the ground plane point given by the projection of the optical axis. From Equation (10), the vertical height at this point $h(i=i_0)$ may be related to the look-down angle as follows

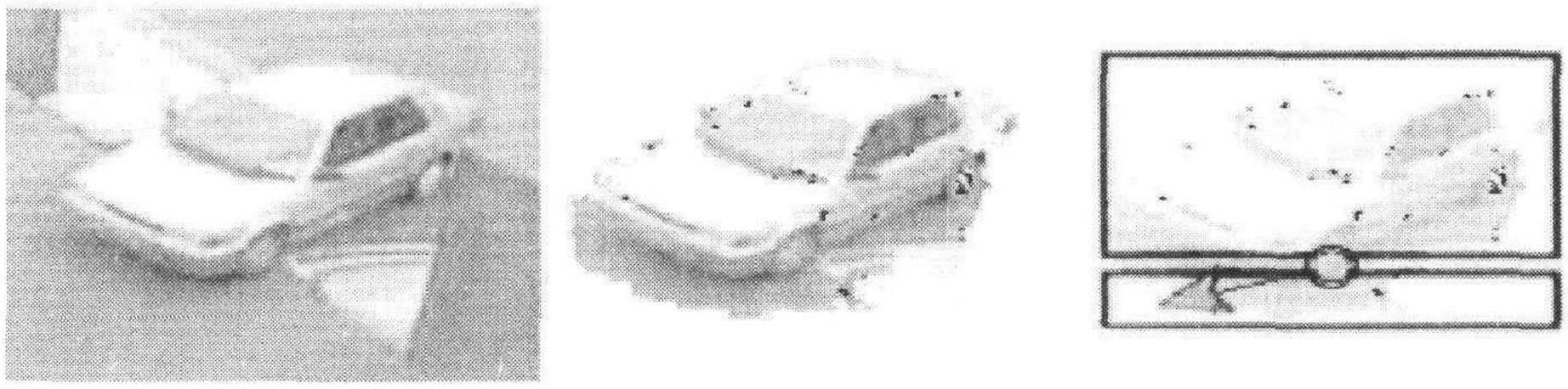
$$\frac{h}{f} = \frac{H\cos\theta\sin\theta}{L - H\cos^2\theta} \tag{12}$$

An estimate of the height h may also be generated from the learnt linear projected height model *i. e.* $h(i_0)=\alpha_x\gamma(i_0-i_h)$. Combining this with Equations (11) and (12), the following expressions for the camera parameters θ and α_y^f may be derived

$$\sin^2\theta = \frac{\gamma}{H}\frac{L-H}{1-\gamma}, \quad \alpha_y^f = \frac{\cot\theta}{(i_0 - i_h)} \tag{13}$$

3 Model-based tracking

In this section, the projected height concept is employed to define simple yet highly effective bounding box appearance models for the principle object types within a surveillance scene. The representation is composed of two vertically adjacent connected bounding boxes - the object component and base component. The base is the large number of background pixels beneath an object and the shadow regions which are typically segmented with the object pixels themselves. The object component is defined by i) the vertical extent of the object - the height model, ii) the horizontal extent of the object - the width model, and iii) the vertical extent of the base region - the base model. These models, as illustrated in Fig. 4, are defined relative to the 2D position of the object - the 2D projection of the position of the object on the ground plane. Three different models are currently used corresponding to each of the principle vehicles types Ψ in the set $\Psi=\{Person, Vehicle, Large\ Vehicle\}$. As with the ground-plane auto-calibration, the parameters for each of these models must be computed in a learning procedure.



(a) Typical object (b) Detected pixels (c) Object model

Fig. 4 Modelling detected events: Images show

$$\begin{aligned} \mu &= \Gamma^{\Psi}(i - i_h) \\ \omega &= \Omega^{\Psi}(\theta)(i - i_h) \\ \beta &= \beta^{\Psi}(\kappa)\Gamma^{\Psi}(i - i_h) \end{aligned} \tag{14}$$

The Height Model. The expected pixel height μ (see Equation (14)) varies linearly with vertical image position i. Different height models $\Gamma^{\Psi}, \psi \in \Psi$ must be defined for each type of object ψ - see Fig. 3(b). A further assumption is made that the projected height of vehicles does not depend on the orientation of the object.

The Width Model. For vehicular objects, the projected pixel width ω varies as a function of depth (and hence varies linearly with position i) but also varies as a function of the 3D orientation of the object. The 3D orientation of a moving vehicle is correlated with the direction of its visual motion. This relationship can be clearly demonstrated in Fig. 3(a) which plots 2D width (normalised by vertical height) against the visual motion direction θ for a large set of detected regions. Thus the projected width of an event is a function of both i and the direction θ.

The Base Model. The vertical extent of the base again varies linearly with the vertical image position (Fig. 4(c)). In dull weather conditions, this base area is usually a small fraction. However in bright weather conditions, this base area can become significantly larger. Currently, the base model parameter β^{Ψ} is manually set as a proportion of the height model. Ideally some environmental illumination parameter κ would select the appropriate ratio.

4 Registering multiple cameras

In Section 2.1 the positions and velocity of objects tracked in each field of view were back-projected onto a local reference frame set on the ground plane. The transformation between cameras is unknown but it can be easily calculated if the correspondences between object positions are known between views[9]. In our auto-calibration scenario, we cannot assume that the correspondences are known. Further, while we assume the availability of object positions with associated velocity vectors, no temporal association is assumed.

The Hough transform approach[10] has been adopted to recover the inter-camera transformation by taking advantage of the fact that the ground plane coordinate systems of temporally synchronised observations of the same 3D object are related by a simple rotation ψ and translation T transformation.

$$\begin{aligned} \boldsymbol{X}' &= R(\psi)\boldsymbol{X} + \boldsymbol{T} \\ \boldsymbol{V}' &= R(\psi)\boldsymbol{V} \end{aligned} \tag{15}$$

where $\boldsymbol{X},\boldsymbol{V}$ and $\boldsymbol{X}',\boldsymbol{V}'$ are positional and velocity observations measured in the local GPCS of two cameras C and C' respectively. Note that the velocity estimates are computed from the partial derivatives of Equation (2) respect to image coordinates, and the 2D tracker image position (i,j) and visual velocity (u,v) estimates *i. e.*

$$V_X = \frac{\partial X}{\partial i}u + \frac{\partial X}{\partial j}v, \quad V_Y = \frac{\partial Y}{\partial i}u + \frac{\partial Y}{\partial j}v \tag{16}$$

In every frame interval, each camera outputs a set of measurements about all objects located in its field of view. As object correspondences are unknown, every pair of observations from each of the cameras must be used to generate a candidate estimate of the transformation. Given a pair observations $\boldsymbol{X}'_{t,i},\boldsymbol{V}'_{t,i}$ and $\boldsymbol{X}_{t,j},\boldsymbol{V}_{t,j}$ at time t from cameras C' and C respectively, transformation estimates may be defined as

$$\begin{aligned} \cos\psi_{t,i,j} &= \hat{\boldsymbol{V}}'_{t,i} \cdot \hat{\boldsymbol{V}}_{t,j} \\ \boldsymbol{T}_{t,i,j} &= \boldsymbol{X}'_{t,i} - R(\psi_{t,i,j})\boldsymbol{X}_{t,j} \end{aligned} \tag{17}$$

where $\hat{\boldsymbol{V}}$ is the unit vector in the direction of $\boldsymbol{V}$. If Λ_t and Λ'_t are the sets of observations in frame t for cameras C and C' respectively, then the set of all observations

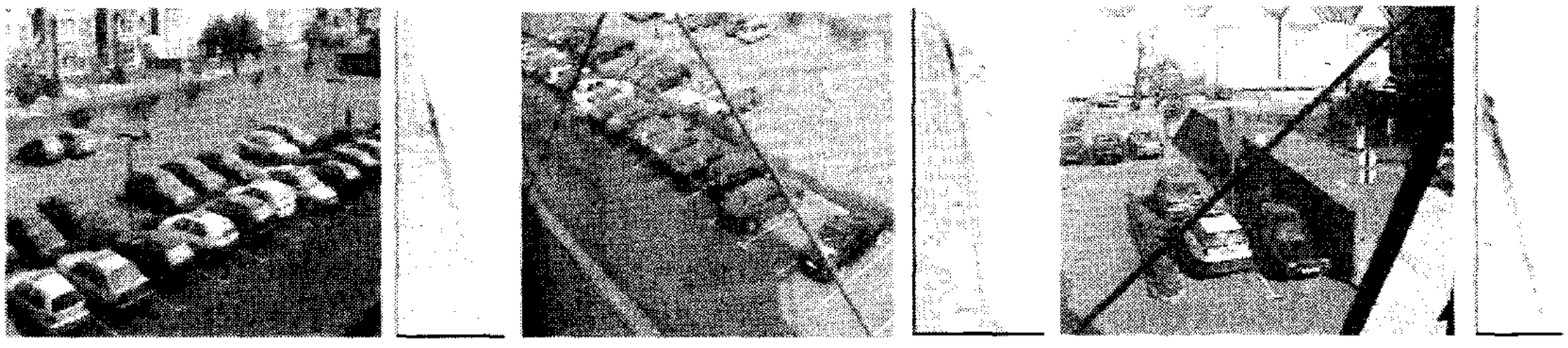
$$\{\psi_{\tau,i,j},\boldsymbol{T}_{\tau,i,j};\forall\, i \in \Lambda'_\tau, \forall\, j \in \Lambda_\tau, \forall\, \tau \leqslant t\} \tag{18}$$

should ideally exhibit a distinct cluster of estimates around the true solution ψ, T within an noise floor of uncorrelated false estimates generated by incorrectly corresponded observation pairs and noise observations. To detect this cluster, the space could be tessellated into bins and a Hough transform technique applied to locate the maximum that correspond to the optimal transformation parameters. However, the range of translation values required is difficult to predict a priori. Therefore to avoid the storage of the problems such a voting

strategy introduces, a robust clustering approach is adopted. The expectation-maximisation mixture of Gaussian technique was implemented and adapted to iteratively perform the cluster analysis on the incoming stream of transform estimates. The clustering process continually reports the most likely transformations between cameras.

5 Results

In the following sections we evaluate the three stages of the overall approach separately. In Section 5.1 the accuracy of the recovery of the local ground plane is tested by comparing the actual and estimated look-down angles. The Tsai calibration results performed on the PETS2001[1)] were not particularly accurate at estimating the camera height and look-down angle. Consequently the evaluation was performed on the three local installations illustrated in Figures 5(a),(b) and (c).



(a) Camera 1 (b) Camera 2 (c) Camera 3

Fig. 5 The TEST installation viewpoint figures (a), (b) and (c) with their corresponding image height vs image vertical position histograms

In Section 5.2 the Ground Plane Tracker (GPT) is compared to the Image Plane Tracker (IPT) and further summarised. Section 5.3 illustrates the process of camera ground plane registration, and evaluates the accuracy of the camera registration results on these and the PETS datasets.

5.1 Image to ground plane calibration

The test installations illustrated in Figure 5 involve different types of camera placed at different heights overlooking a common car park scene. The car park has been surveyed to generated real-world ground plane positions in a common coordinate system. These points have been selected to ensure that each camera has ten well distributed points in the image plane. The convex hull of these points contains most of the car park and over fifty percent of the visual plane. The real look-down angles and camera heights have been established using surveying equipment from the ground plane projection of the correct optical axes.

As described in Section 2.3 the projected height model for each camera can be recovered by accumulating in a height versus vertical image position space and fitting a straight line to the resultant histogram. Results for Cameras 1, 2 and 3 are shown in Figures 5 (a), (b) and (c). In conjunction with the measured height of the cameras above the ground plane, the parameters of these models can be used to derive the extrinsic and some of the intrinsic camera parameters - see Equation (13). To compare the accuracy of the proposed method, the ground truth data, the traditional Tsai[8] technique results and the measurements generated by our approach are tabulated in Table 1. In all cases, the accuracy of the Tsai method and our own is comparable, with the shallow angle of view of Scene 3 being the most problematic. We employed the Tsai results to confirm that the camera had no significant roll *i. e.* rotation around the optical axis - typically less than 4°. The method proposed in this work accurately located the look-down angle although care had to

1) The PETS2001 datasets (visualsurveillance.org) are problematic as they contain so few tightly distributed calibration points.

be taken to correctly fit the linear model to the projected height ridge of the histogram.

Table 1 Look-down angle results. For clarity the look-down angle has been redefined as $\pi/2-\theta$ defining the angle of intersection between ground plane and optical axis

Test Installation	Correct Height (m)	Tsai Height (m)	HRE γ	Horizon v_h	Correct Angle (°)	Tsai Angle (°)	Our Approach (°)
Camera 1	9.1	9.9	0.195	−22.3	16	16.7	15.5
Camera 2	13.9	15.4	0.109	−174	24.3	24.5	23.3
Camera 3	6.7	5.7	0.255	17	13.5	7.7	11.7

5.2 Model based tracking

The Ground Plane Tracker (GPT) embeds the mechanisms introduced in this paper within a standard tracking framework, and is compared against a standard 2D tracker - the Image Plane Tracker (IPT). Both mechanisms employ a Kalman filter model whose observation and dynamic noise models are learnt directly from the data. The two methods are summarised in Table 2 below. Data association is performed by searching predicted bounding boxes for union of overlapping moving regions whose area is greater than 10% of bounding box area. Model instances are instantiated from unassigned moving regions[10] whose areas are greater than some common threshold - 10 pixels (in quarter-size PAL frames). In neither case is any additional appearance matching implemented to improve data association. Observation position error is defined as deviation from predicted object dimension. Each object has a time-to-live counter (TTL) defined as $\min(TTL, 10)$ which is incremented if inter-frame match recovered, and decremented if no match recovered with object deleted when $TTL<0$.

Table 2 Implementation details of standard and proposed tracking algorithms

Algorithm	Image Plane Tracker	Ground Plane Tracker
Measurement	x, y image pixels	X, Y ground plane-Eq. (2)
Motion Model	First-order $x, y, \dot{x}, \dot{y}$	Second-order $X, Y, \dot{X}, \dot{Y}, \ddot{X}, Y$
Appearance Model	First-order Kalman filter on bounding box dimensions $h, w, \dot{h}, \dot{w}$	Position and velocity constrained bounding box model of Eq. (14)

To compare the different approaches a tracking error is defined as the number of track failures per 1000 track frames. A track failure occurs when the tracking identity of any ground truth object changes. Track frames are the total number of object appearances for all tracks in a sequence. The experiment is run on three different datasets-see Table 3: the PETS 2001 Dataset 1 (an occlusion rich dataset of distant objects in good lighting conditions), the Kingston Car Park Dataset (although relatively free of non-static occlusions, objects exhibit considerable motion variation against background undergoing frequent and severe lighting variations caused by intermittent direct and reflected sunshine), and the Football Dataset (large number of objects undergoing correlated and rapidly changing motions). Note that the tracking results reflect the challenging nature of the Kingston data sets and, in particular, the Football Dataset. Nonetheless, the proposed tracker outperforms the traditional tracker which is easily misled. Greater insight into the problems of trackers can be gained by determining the nature and frequency[1] (% of frames) of the failure modes—see Table 4.

Table 3 Tracking error

Tracker	PETS	DIRC	Football
IPT	3.2	1.5	49
GPT	1.9	1.1	11

1) Frequency will be highly dependent on dataset.

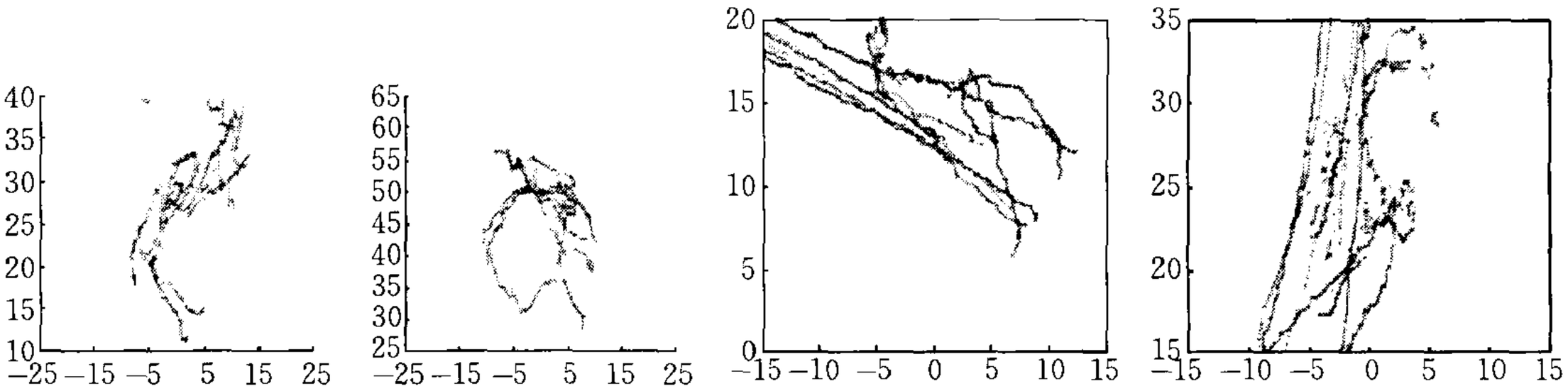
Table 4 Tracking error

Failure	Description of data association failure	Failure	
		IPT(%)	GPT(%)
Fragmentation	Unexpected small displaced observation	9	2
Static occlusion	Unexpected small displaced observation	23	10
Object occlusion	Unexpectedly large observation	36	34
Motion model	Motion model too constraining	21	34
Stationary object	Object merges into background	11	20

Both trackers loose track of objects that are stationary for several seconds-determined by a TTL parameter. However the principal weakness of the traditional tracker is when dealing with situations where i) fragmentation or static occlusion processes shrink the search window with consequent failure to locate validating observations, and ii) occlusions which widen the search window causing the tracker to be deflected by the occluding object. These problems are more likely in situations where the trajectory deviates from the assumed motion model.

5.3 Multi-camera calibration

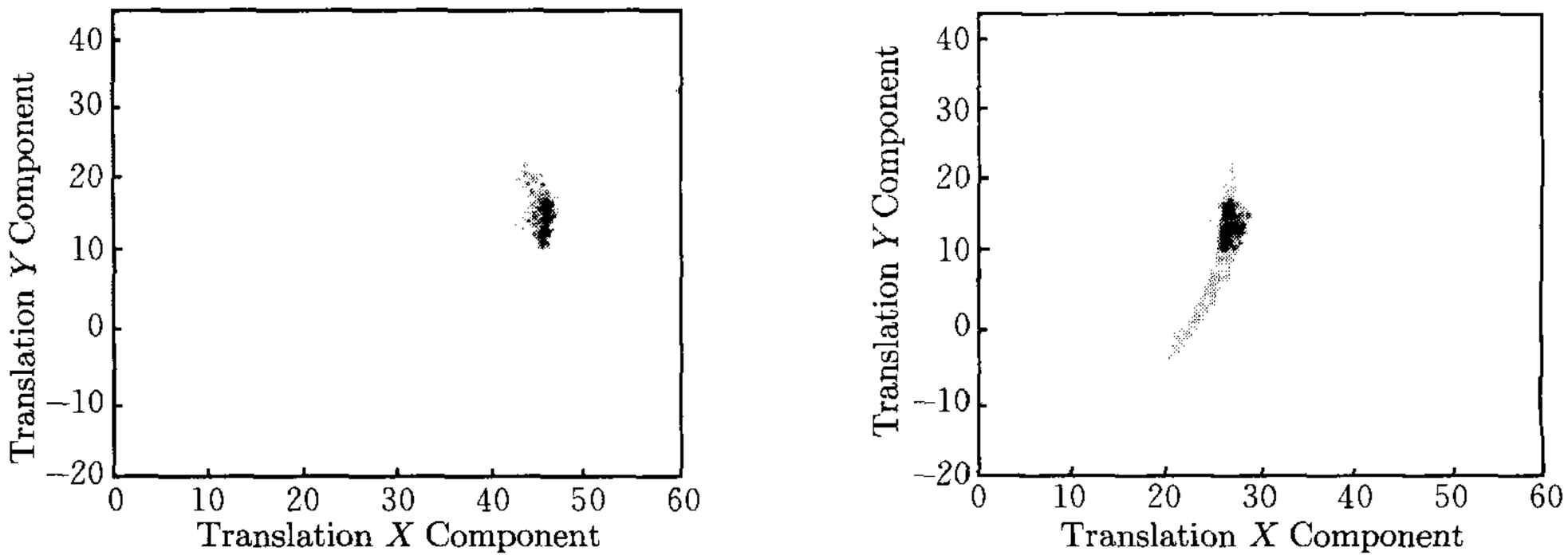
Fig. 6 plots the tracked object trajectories recovered from our motion detection and tracking software[11] and projected onto the local ground plane of each camera.



(a) TEST Camera 1 (b) TEST Camera 2 (c) PETS Camera 1 (d) PETS Camera 2

Fig. 6 Projected trajectories: note only a roughly contemporaneous set of trajectories are plotted

These observations are used to build the rotation and translation Hough space described by Equations (17). The populated space and dominant cluster are shown in Figures 7(a) and (b) for the TEST and PETS datasets respectively. Note that these peaks are robustly recovered from an extensive noise floor. Produced by computing registration estimates for every pair of trajectory observations, this floor arises from the need to avoid the prior establishment of observation correspondences.



(a) TEST (b) PETS

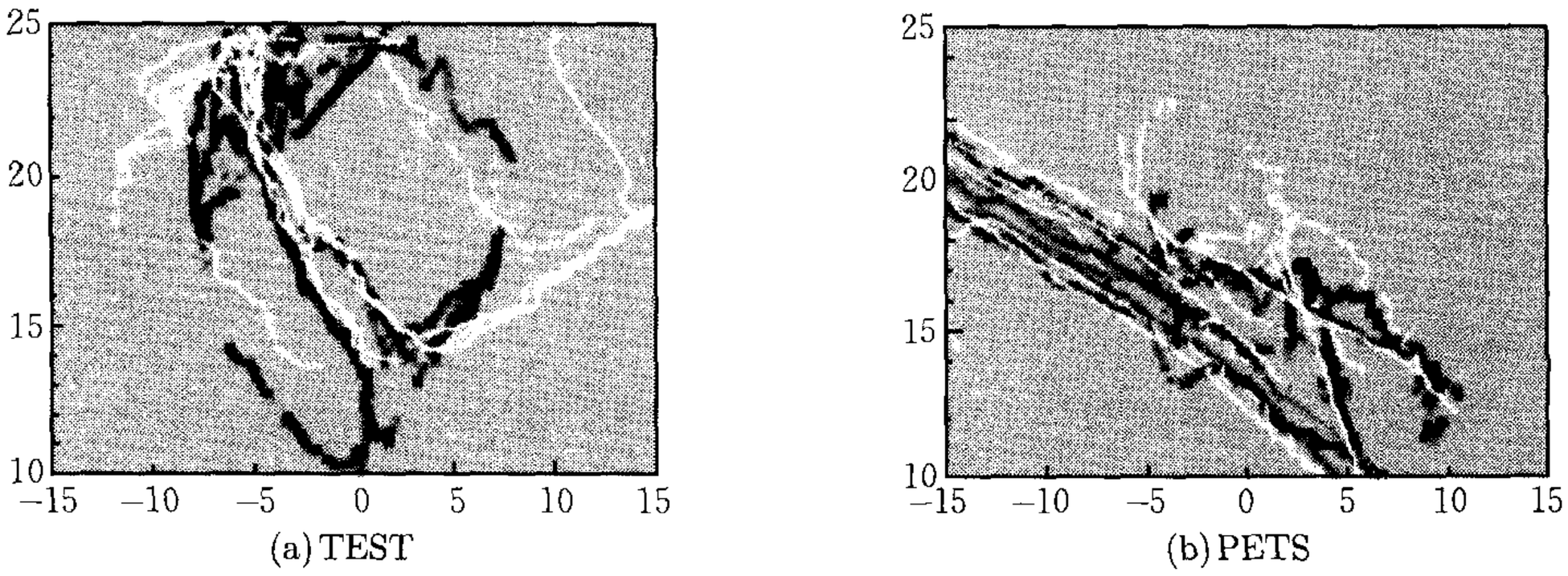
Fig. 7 Locating the maximal cluster in transform space

The accuracy of the technique may be judged as before by comparing the registration results of the method with the equivalent data supplied by the Tsai calibration method (and the ground truth for the TEST datasets). Table 5 plots the rotation angle ψ (in degrees) and distance $|\boldsymbol{T}|$ (in metres) between the origins of the two local GPCS for the TEST and PETS datasets.

Table 5 Registration results

Datasets	Measurements			
	Tsai		Proposed	
	$\psi(^\circ)$	$\lvert\boldsymbol{T}\rvert$	$\psi(^\circ)$	$\lvert\boldsymbol{T}\rvert$
TEST	75	57.2	81	53.5
PETS	76	27.9	70	29.5

Despite the poor accuracy associated with off-ground-plane estimates, the accuracy of the ground plane projections are in agreement with the correct values surveyed in the DIRC datasets. Thus only the Tsai results are quoted in Table 5. The recovered values for the rotation angle ψ and distance $|\boldsymbol{T}|$ are used to rotate and translate the data into a single coordinate system (that of the second camera). The overlapped trajectories are displayed in Fig. 8.



(a) TEST (b) PETS

Fig. 8 Overlaying trajectories

While not in perfect alignment, the accuracy appears sufficient to establish the correspondence of any new objects that enter the scene. Any lack of alignment arises from a number of sources: i) any existing roll angle on either camera; ii) inaccuracies in estimation of the look-down angle and intrinsic parameters of either camera; and iii) view-dependent positional bias in trajectories. The presented results are based on the location of the foot of an object on the ground plane. This position has demonstrated a strong viewpoint dependency when applied to car objects or person objects in the presence of shadows. A more consistent vertically weight centroid position will increase the degree of alignment.

6 Conclusions

A central objective of this work focuses on the development of learning techniques for use in plug-and-play visual surveillance multi-camera systems. Many camera calibration techniques exist, however most of them require the assistance of an expert to tune a set of parameters. The underlying strategy is to develop a suite of algorithms that could be installed by non-technical personnel, and as much as possible based on self-adjusting techniques that learn how to adapt to the camera set-up, the environmental changes and possibly to weather conditions. This paper presents a novel camera calibration approach, based on two separate stages.

In the first stage a linear model of the projected height of objects in the scene is used in conjunction with world knowledge about the average person height and the height of each camera to recover the image-plane to local-ground-plane transformation of each camera. In the second stage, a clustering technique (based on expectation-maximisation) is then used to recover the transformation between these local ground planes. A comparison between the proposed technique and the standard approach of Tsai was carried out. Results for both techniques, evaluated with ground truth measures, show that the accuracy of the proposed approach is similar to Tsai's approach.

Although a more detailed evaluation is required, the presented preliminarily results demonstrate that approach generates sufficient accuracy to enable trajectory data to be fused within a common ground plane coordinate system between each pair of cameras. In particular, to robustly support the plug and play the sensitivity of the approach to violations in the assumptions of i) projected height linearity and ii) zero-roll angle must be investigated. Finally, the method had to be tested on a new data set rather than the PETS2001 images as the lack of calibration points makes the recover of accurate camera height problematic.

References

1 Stauffer C, Grimson W E L. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(8):747~757

2 Bar-Shalom Y, Fortmann T. Tracking and data association. In:Mathematics in Science and Engineering. Academic Press, 1988

3 Siebel N T, Maybank S J. Real-time tracking of pedestrians and vehicles. In:Proceedings of the 2nd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Hawaii:IEEE Press, 2001

4 Haritaoglu I, Harwood D, Davis L S. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(8):809~830

5 Ellis T, Xu M. Object detection and tracking in an open and dynamic world. In:Proceedings of the 2nd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Hawaii:IEEE Press, 2001

6 Fuentes L M, Velastin S A. People tracking in surveillance applications. In:Proceedings of the 2nd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Hawaii:IEEE Press, 2001

7 Cupillard F, Bremond F, Thonnat M. Tracking groups of people for video surveillance. In:Proceedings of the 2nd European Workshop on Advanced Videobased Surveillance Systems, UK: Kingston, 2001. 88~100

8 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*, 1987,**3**(4):323~344

9 Haralick R M, Joo H. 2D-3D pose estimation. In:Proceedings of the International Conference on Pattern Recognition, 1988. 385~391

10 Ballard D H, Brown C M. Computer Vision. New Jersey:Prentice-Hall, Inc., 1982

11 Orwell J, Remagnino P, Jones G A. From connected components to object sequences. In: Proceedings of the 1st IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2000. 72~79

John-Paul Renno Received his bachelor degree in Electronic Engineering with Computing (2000) from Kingston University, Surrey, UK. Since then he has been a researcher in Computer Vision studying for his Ph. D. degree in Surveillance systems. The principal interests of the project are change detection and the tracking of objects in image sequences captured from CCTV surveillance systems.

Paolo Remagnino Joined the Digital Imaging Research Centre at Kingston University in September 1998. As an active researcher in the field of Machine Learning and Computer Vision, he received Ph. D. degree in Computer Vision from the University of Surrey (Guildford, UK), and has over 12 years research experience in Artificial Intelligence methodologies. He is currently project manager of three British Council projects involving international collaboration with other EU Member states (Spain, Italy and France) all related to machine learning techniques applied to computer vision applications. Dr Remagnino is a member of the scientific committee of the International Association of Science and Technology for Development (IASTED), and is currently a Guest Editor for a Special Issue of IEEE Transactions on Systems, Man and Cybernetics Part A: Systems and Humans on Ambient Intelligence.

Graeme Jones Director of the Digital Imaging Research Centre at Kingston University. He has a Ph. D. in computer vision from Kings College London, and has over 15 years experience in image and video sequence analysis, intelligent systems and multimedia data communications, and has managed a number of industrial and government funded projects in these areas. Recently he been responsible for video/image analysis projects supported by the film and special effects industry (Computer Film Company Ltd., UK and Dynamic Digital Depth Pty. Ltd., Australia) and by the video security in

dustry (Primary Image Vision Systems Ltd.). He is currently co-investigator on the EU INMOVE (IST-2001-37422) project developing an expandable set of software tools enabling the provision of a new range of intelligent video based services to end users in various mobile/wireless networks. In 2000, Dr. Jones chaired the British Machine Vision Workshop on Visual Surveillance and was co-chair of the IAPR Workshop on Advanced Video-based Surveillance Systems in 2001.

针对自校准地平面的学习监控跟踪模型

J. Renno P. Remagnino G. A. Jones

(*Digital Imaging Research Centre*, *Kingston University*, *Kingston*, 英国)

(E-mail: {j. renno, p. remagnino, g. jones}@kingston. ac. uk)

摘 要 提出了一种新的多摄像机视觉监控系统的信息融合方法. 信息融合在两个阶段进行. 首先,根据相互独立的 Cartesian 参考坐标系统(设置在地平面上),对各个摄像机进行标定. 然后,把所有的坐标系变换到一个坐标系统中. 在视觉监控应用中,因为摄像机自定标和视觉数据配准技术将使监控设施安置变得更加容易,从而可以为公共场合发展更加适用的视觉监控工具. 在解决监控数据的不完整性和不确定性方面,机器学习方法具有很好的效果.

关键词 视觉监控,机器学习,数据融合,摄像机标定

中图分类号 TP391.41

(Continued on Page 465)