1 Leeds People Tracker

1.1 Introduction

In this section we give a brief overview of the pedestrian tracking system developed in Leeds.

We have found that a flexible 2D prior model of silouhette shape is an effective approach to recognising and tracking walking pedestrians in image sequences. This is in contrast to more conventional 3D hand-crafted part-based models (e.g. Gavrilla and Davis [1], Rohr [2], Hogg [3]).

A prior shape model incorporates useful constraints on the apparent shape of a pedestrian silhouette that allows the system to cope with missing information due to image noise, background clutter and partial occlusions. A flexible (or deformable) model is required to model the apparent change in shape due to pose (i.e. position of limbs etc) and viewpoint relative to the camera.

An important feature of this approach is that the model is learned *automatically* using training image sequences containing the object of interest. This allows the system to be trained up on a particular viewpoint to learn specific shape constraints appropriate to the given object in the given scenario. A further benefit of this approach is the very low dimensionality of the model and the linear nature of the transformation from model space to image space.

A statistically based tracking system is used to estimate model parameters for each moving pedestrian in the scene. The method used relies on a Kalman filter mechanism which proves computationally efficient allowing real-time tracking performance for several moving pedestrians.

1.2 Model generation

The shape model is derived from a set of representative training shapes. The method is described in previous work [4, 5]. In summary, the training shapes are automatically segmented using conventional image subtraction (alternatively a "snake", as described by Kass [6], could be applied to high quality training images). The boundary of each segemented training shape is traced automatically and then represented in a consistent manner using a cubic B-spline contour with N control points space equally around the boundary (see [4] for details).

This consistent shape parametrisation allows the spline control points representing each training shape to be treated in a similar manner to the "landmark" features of Cootes *et al* [7] and Bookstein [8]. The shapes are

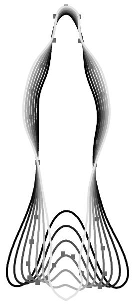


Figure 1: Principal shape mode

aligned to a reference frame and Principal Component Analysis is performed on the spline data (using an appopriate distance metric for this shape-vector space [5]).

The resulting prior shape model consists of the mean shape and a relatively small set of "modes of variation" (eigenshapes) that encapsulate a high proportion of the variability observed in the training data. The model can be made more compact by iteratively reparametrising the training shapes using the current shape model (see Baumberg and Hogg [5]). The resulting principal mode of variation is illustrated in figure 1

1.3 Kalman based image fitting and tracking

Each pedestrian object hypothesis in the scene can be represented by the object centre image coordinates (o_x, o_y), the scale and orientation s and θ and a small set of shape parameters b_i which are weights for the associated eigenshape basis vectors.

An iterated Kalman filter provides a useful framework for parameter estimation on a single static image. The method summarised here has been described in previous work [9, 5, 10] and is based on the earlier work of Blake [11]. The object state parameters are grouped into the position parameters (and the velocity), the alignment parameters (rotation and scale) and each shape parameter. For computational speed, each group is filtered independently.

1.3.1 Initialisation

A Kalman filter is maintained for each shape parameter. The filter holds an estimate of the shape parameter, $\hat{b}_i$ and the associated variance for the estimate, σ_i. The shape parameters are initialised to zero (i.e the mean shape) and the associated variance set to the variance of the parameter over the training set. The position, rotation and scale parameters are initialised using a motion-based event detector.

1.3.2 Measurement step

At each iteration, n regularly spaced sample points are calculated along the estimated B-spline contour. Measurements are made by searching for suitable features (such as large changes in intensity) along an optimal search line related to the contour normal direction (see [10]).

The search region is determined by searching along each line within an uncertainty ellipse constructed from the positional covariance. The search window size, is chosen so that the measured feature lies within a Mahalanobis distance of 2 standard deviations from the expected position.

Each point measurement has an associated pointwise measurement covariance matrix. Simplifying matters, the "variance" of each measurement is set proportional to the square size of the search window at that point and the total number of measurements that will be made (assuming a fixed number of iterations). However, if no significant feature is found within the search region or the search region is less than one pixel, the "variance" is set to be infinite.

1.3.3 Filter update

Using the standard Kalman filter equations the measurements are combined to update the parameter estimates and their uncertainties. The process is then iterated using a fixed number of iterations or until convergence. Between image frames a statistical model is used to estimate the new object parameters and their associated uncertainties. Each shape parameter is assumed constant with a random perturbation term proportional to the variance of the shape parameter over the training set. Alternatively a spatiotemporal model can be learned to give better predictions for walking pedestrians [12]. The object position is assumed to have a constant velocity with a random acceleration component.

1.3.4 Ground plane filtering

When the ground plane is known, this can be used to constrain the filtering process. The image-centred height and position parameters can be mapped to world-centred ground plane position and world height parameters. The world-centred parameters are filtered in a similar manner to the image-centred parameters and used to provide more accurate estimates for image poisition and image height. The associated covariance matrices can be projected between world and image frame by linearising the camera projection transformation using the current estimate of object depth.

1.4 High level control

1.4.1 Hypothesis generation

A simple motion-based event detector is used to detect new untracked objects. The detector uses spatial and temporal subsampling to reduce the computation load. Pixelwise image subtraction with a dynamically updated "background" image is used. Thresholding is performed and the silouhette for each of the current hypotheses masked from this binary image. The remaining significantly sized motion regions are used to generate new hypotheses.

At the hypothesis generation stage a person is considered to be the average world height and to be vertical in the image. This allows two hypotheses to be generated for each motion region.

1. The top line of the region bounding box corresponds to the top of a person's head

2. The bottom line of the region bounding box corresponds to the bottom of a person's feet

Each hypothesis corresponds to a different ground plane object location. Image evidence for both (possibly) conflicting hypotheses is examined and the less likely hypothesis rejected.

1.4.2 Validation

The image fitting process described previously is used to provide a "fitness" measure for the current hypothesis. The fitness measure is the percentage of the contour that is "locked" onto a significant image feature. If the fitness for a hypothesis is consistenly high over several frames then the hypothesis is accepted. Otherwise the hypothesis is rejected and no longer examined.

1.4.3 Track termination

Once a hypothesis is accepted as a moving person the object is continually tracked until the ground plane positional uncertainty becomes too large and the object is lost. This can occur when a person walks off the camera image or enters a building or car and is totally occluded for a significant length of time.

2 Interaction with the car tracker

2.1 Hypothesis generation

The motion event detector uses the information from both tracking systems to mask out pixels where there is motion from currently tracked objects.

The remaining regions of significant motion are then examined and region arbitration performed using the size of the region. In this application large regions usually correspond to cars although future implementations may use more complex schemes. For instance, the motion event may generate two confliciting hypotheses that the event corresponds to both a car and a person and then reject the hypothesis with less supporting image evidence.

One further complication is worth noting. In a car park scenerio cars are often parked for a long period of time. In order to detect motion in front of these cars, we allow the car to become incorporated into the "background" image. This is achieved easily by dynamically updating the background image using an approximation to a pixelwise median filter over time (see for example McFarlane and Schofield [13]).

2.2 Occlusion handling

Partial occlusion has two effects on the tracking process.

1. Distortion of measurements: The image fitting scheme is driven by image measurements such as local edge and motion based features. Partial occlusion (or near occlusion) by a second moving object will often distort these measurements because a false significant feature may be found that corresponds to the wrong object.

2. Hiding features: If a person walks behind a car then the image features may be obscured and no significant feature found within a sample point search window. This increases the total variance of the measurement process.

The second effect is not undesirable because where there is partial occlusion the confidence in the measurements should decrease.

Occlusion is handled by the tracker by ignoring measurements where there is likely to be occlusion. For each tracked person an occlusion mask

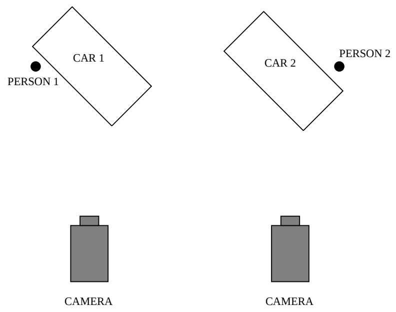


Figure 2: Example where simple depth ordering fails

image is generated. If the search window for a given sample point overlaps a pixel which is set (where there is likely occlusion), then the measurement for that sample point is ignored.

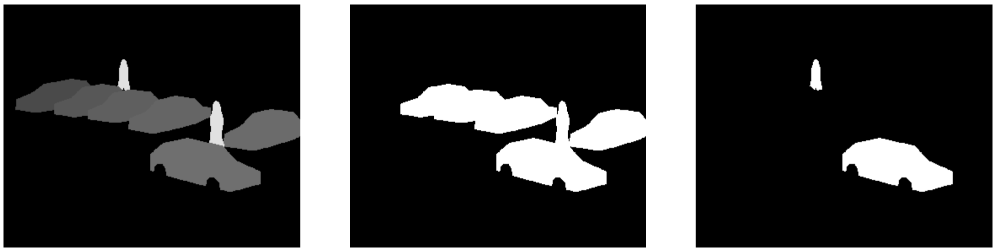
2.3 Generating a person centred occlusion mask

In general, for each tracked object *in front* of a person the associated silouhette region in the occlusion mask should be set. Note that we need a "footprint" map to decide wether a person's ground plane position (considered as a point) is in front or behind each car's ground plane bounding box. A simple depth ordered scheme such as that used by Koller *et al* [14] would not be sufficient to disambiguate all cases. For instance, in fig 2 the depth of Person 1 and Person 2 are the same (relative to Car 1 and Car 2) but only Person 2 is occluded. A car is considered *in front* of a person if the ray from the person to the camera crosses the car bounding box on the ground plane.

If motion based features are used (i.e. using local background image subtraction), the problem becomes more involved. In such cases, moving objects mutually occlude one another (irrespective of depth). However, stationary objects that have been dynamically incorporated into the background image (e.g. a parked car) occlude only those objects behind them. Things become more complex where there are several moving objects and stationary objects (although the implementation turns out to be relatively simple).

2.3.1 Validation

The fitness measure described previously can be easily extended to account for partial occlusion. The extended fitness measure is now defined as the percentage "lock" over the number of sample points where no occlusion was



(a) **(b)** **(c)**

Figure 3: Example occlusion masks

observed. However a further constraint is required to validate a hypothesis. If the estimated contour is highly occluded then a few spurious image features may result in an inappropriately high fitness measure. To account for this, an additional constraint is required. A hypothesis is only accepted if a significant fraction of the contour is unoccluded.

3 Experimental Results

3.1 Occlusion handling

Figure 3 shows the occlusion masks for two tracked pedestrians (where motion-based image features are used). The image frame is visualised in figure 3 (a). The occlusion masks for the furthest and nearest pedestrian are shown in figures 3 (b) and (c) respectively. Note that in each case the second pedestrian is included in the occlusion mask. This is because moving objects mutually occlude one another when motion-based image features are used. The nearest car is also moving and occludes both pedestrians. The remaining cars are stationary and have been incorporated into the background image. These staionary cars only occlude pedestrians that are behind them.

Figure 4 shows the image feature search lines that are unoccluded for a particular pedestrian hypothesis. The tracker is able to recognise and track this object even though it is initially partially occluded.

References

[1] D M Gavrilla and L S Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *IEEE Computer Vision and Pattern*

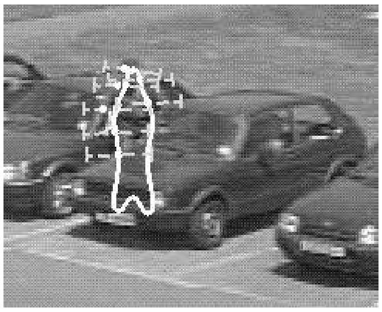


Figure 4: Image fitting with partial occlusion

Recognition, 1996.

[2] K Rohr. Incremental recognition of pedestrians from image sequences. *Computer Vision and Pattern Recognition*, pages 8–13, 1993.

[3] D Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[4] A Baumberg and D Hogg. Learning flexible models from image sequences. In *European Conference on Computer Vision*, volume 1, pages 299–308, May 1994.

[5] A Baumberg and D Hogg. An adaptive eigenshape model. In D Pycock, editor, *British Machine Vision Conference*, volume 1, pages 87–96. BMVA, September 1995.

[6] M Kass, A Witkin, and D Terzopoulos. Snakes: Active contour models. In *First International Conference on Computer Vision*, pages 259–268, 1987.

[7] T F Cootes, C J Taylor, A Lanitis, D H Cooper, and J Graham. Building and using flexible models incorporating grey-level information. In *International Conference on Computer Vision*, pages 242–246, May 1993.

[8] F Bookstein. *Morphometric Tools for Landmark Data.* Cambridge University Press, 1991.

[9] A Baumberg and D Hogg. An efficient method for contour tracking using active shape models. In IEEE Computer Society Press, editor,

IEEE Workshop on Motion of Non-rigid and Articulated Objects, pages 194–199, November 1994.

[10] A Baumberg. Hierarchical shape fitting using an iterated linear filter. In *British Machine Vision Conference*, volume 1, pages 313–323. BMVA, September 1996.

[11] A Blake, R Curwen, and A Zisserman. A framework for spatio-temporal control in the tracking of visual contours. *International Journal of Computer Vision*, 1993.

[12] A Baumberg and D Hogg. Generating spatiotemporal models from training examples. *Image and Vision Computing*, June 1996. in press.

[13] N J B McFarlane and C P Schofield. Segmentation and tracking of piglets in images. *Machine Vision and Appications*, 8:187–193, 1995.

[14] D Koller, J Weber, and J Malik. Robust multiple car tracking with occlusion reasoning. In *European Conference on Computer Vision*, volume 1, pages 189–196, May 1994.