Person Tracking, Pursuit & Interception by Mobile Robot

Punarjay Chakravarty, David Rawlinson, Ray Jarvis
Intelligent Robotics Research Centre, Monash University
{ punarjay.chakravarty | david.rawlinson | ray.jarvis }@eng.monash.edu.au

Abstract

We present a method of tracking a human target whilst navigating in an unknown environment using a combination of visual and laser range sensing. The target's legs detected by laser initialize both the pursuit process and modelling of target visual appearance. Intermittent recognition & tracking is possible using the learned visual model, as would likely occur in real pursuit scenarios when the target is temporarily obscured by obstacles. Use of range sensors and coordination of visual target tracking and obstacle avoidance behaviours allow operation in realistic environments. The system is applicable to both security and domestic assistance tasks.

1 Introduction

1.1 Scenario

In order to significantly improve the autonomous capabilities of surveillance and security systems intervention by mobile robot would seem critical. For instance, delivery of specialist tools and sensors to an incident could assist activities as diverse as bomb-disposal and identity verification (e.g. scanning of an ID card).

Cost-effective systems would inevitably combine expensive "specialist" mobile platforms with cheaper and more numerous static devices such as the ubiquitous PTZ cameras. This implies coordination, but ideally without costly installation or configuration requirements, detailed surveys of the environment or a large staff of highly trained operators.

In previous work [Rawlinson, et al., 2004], the authors presented a method by which mobile and fixed cameras can be coordinated without onerous installation, calibration or constraining of the environment. Instead, synchrony of movement between distributed senses permits the construction of inter-modal mappings – for example, the transformation between the camera image and the odometric ground planes. As a result the robot can drive to a target identified on the image plane without non-learned prior knowledge of camera viewpoint or geometry.

This work describes a method by which a mobile agent can model a specified target and then use onboard sensors to pursue (follow). Mobile agents need to be able to work independently using onboard sensors to make up for deficiencies in global surveillance and sensor limitations such as range. Further, it is sometimes difficult to fuse or translate target characteristics between systems and therefore it is important that mobile devices be able to acquire and track targets using onboard sensors.

Autonomous person-following capabilities would also be useful in a domestic setting e.g. a robot could be tasked with carrying items for disabled people or to check on the sick or infirm.

From a technical perspective this paper reports on an algorithm developed for a mobile robot that fuses laser and vision information to acquire, track and follow a person in a cluttered indoor environment. The key features of the system are:

- Visual learning of intruder appearance during an encounter, using fusion of camera and laser data
- Particle-filter based visual tracking of intruder using learnt colour model
- Pursuit of target person while navigating hitherto unknown cluttered environment
- Intermittent pursuit possible – i.e. losing and reacquiring target

1.2 Related Work

The literature includes many successful approaches to human tracking & modelling, many of them being able to discern pose in addition to differentiating between individuals [Haritaoglu, *et al.*, 2006]. Where images of sufficient quality are available, it is also possible to perform detailed analysis such as face-recognition [Turk and Pentland, 1991]. It is possible that a mobile security agent might be able to intercept cooperative human targets and apply some of these techniques using onboard sensors, but for the pursuit or tracking task they are superflous (although in a crowded environment they might be essential). Many human visual tracking schemes are unsuitable for deployment on a mobile device due to reliance on (most commonly) fixed target pose, fixed sensor pose or the presence of much static background. Therefore the citations below focus on approaches proven effective on mobile platforms.

One of the problems with tracking people from a mobile robot is that because the robot itself is moving, it is difficult to segment out the motion of people around it. The following two approaches have tried to model and remove the ego-motion of the robot using vision and laser respectively. After removing the ego-motion of the robot, any motion that remains is considered to be that of humans in the vicinity.

[Jung and Sukhatme, 2004] used a single camera to track people from a moving platform in an outdoor environment. Successive frames were registered using sparse optical flow to remove the image motion due to the ego-motion of the robot. This is followed by frame differencing to extract the motion in the image caused only by moving targets, and particle filtering to track these targets. We have experimented with this approach and found that it works well for image sequences where the person to be followed occupies a relatively small part of the image. Conversely for tracking in enclosed indoor environments, the robot is often in close proximity to the person to be tracked, who occupies a large fraction of its field of view leading to the failure of this method . The other major drawback is that target motion is best picked out when it is in a direction perpendicular to that of the robot, which is rarely the case when a robot is following a person.

[Shulz et.al, 2003] achieved laser-based multiple person tracking from a mobile robot. Laser scan-matching was used to register successive frames, which were then differenced to reveal the positions of moving objects. Multiple targets were then tracked using a particle filter.

[Morate, 2005] uses a combination of laser and vision to track people from a stationary robot. Laser leg segments are detected by using a combination of laser frame-differencing and convexity measurement (searching for arcs where the average angle subtended by arc-extremities lie between two thresholds). This gives a number of candidate leg segments which are checked by panning the camera to them and looking for skin-coloured segments. Skin-colour segmentation is not very successful in our lab environment which has a wooden floor that is confused with skin colour.

[Zhang and Kodagoda, 2005] also used a combination of laser and vision to detect people from a mobile platform. Laser scan-matching is used to register successive frames and identify moving regions. Laser segments define an area of interest in the image which is searched for human templates by a hierarchical template matching strategy. This approach requires the construction of an extensive database of human sillhouettes at different distances from the camera. In addition, the target needs to be visible to the camera in its entirety.

[Treptow, et al., 2006] use an elliptical head-body model to detect people in thermal imagery taken from a mobile robot. Detected people are then tracked using multiple particle filters.

The authors [Chakravarty and Jarvis, 2006] have also used particle filters to track multiple people simultaneously from data obtained from a panoramic camera and a laser range finder mounted on a stationary robot. The panoramic camera gives a 360 degree field of view, but with a reduced resolution. In future, we plan to combine this system with the one presented in this paper, whereby the robot, will be able to track multiple people when stationary, and then pursue one of them.

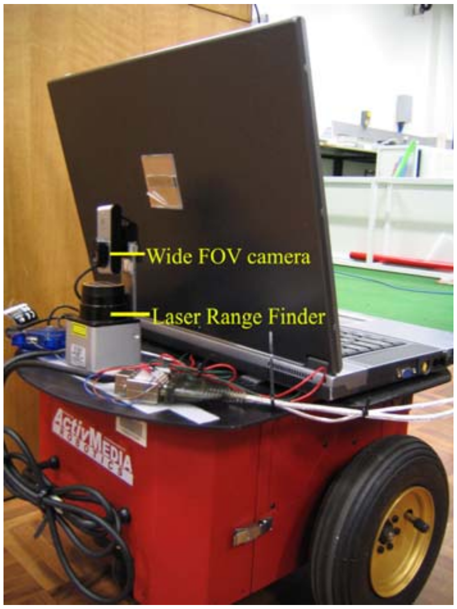

Figure 1 Wide FOV camera and laser range finder mounted on Pioneer robot

1.3 Our approach: Overview

Our system uses a combination of a laser and a 78 degree field of view (FOV) forward-looking camera mounted low on the robot, for person tracking (Figure 1). Its height above the ground means that it rarely sees more than the legs of the person it is tracking. A colour model of the legs is extracted initially using background subtraction, enhanced by a laser-based leg extraction algorithm (scan projected onto image plane). A multi-Gaussian colour model of the target is made and the number of back-projected pixels in each column of the image indicates the presence of legs as peaks. The column in the image is then converted to bearing value by lookup and given to the navigation algorithm.

Since the purpose of the experiment is to demonstrate construction and use of human target models from fused laser/visual data a fast and simple reactive navigation scheme sufficed for controlling movement. However in order to demonstrate the system in a realistic cluttered environment it was necessary to implement obstacle avoidance behaviour, with the additional constraint that the fixed onboard camera must retain visual contact with the target whenever possible. In many situations these two goals are mutually exclusive, since the robot must face away from obstacles to drive around them and has no rear-facing range sensors.

Interestingly, the robot was able to evidence successful pursuit & avoidance behaviour without prior knowledge or mapping of the environment. This suggests that mobile robots could intervene effectively and usefully in security applications – in conjunction with a network of fixed cameras – without costly on-board navigation processing.

2 Algorithm

2.1 Target Detection using vision and laser

Legs appear as arcs in the laser scan. An arc has the property that angles (A1 and A2 in Figure 2) subtended by its extremities are equal.

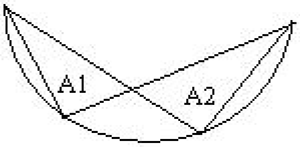

Figure 2 Angles subtended by arc extremities are equal

The laser scan is first divided up into segments based on how far away each point in the scan is from the previous one (Figure 3). Small segments (filtered by the number of laser returns in the segment) are then tested for convexity by measuring the mean and variance of the angles subtended by the segment extremities. If the variance of the angle is within an emprically determined threshold of the mean, then the segment is classified as a leg hypothesis. Leg hypotheses in the laser scan are shown in Figure 4. The leg segments are then projected onto the image plane Figure 5. The laser fails to detect the legs when the person is greater than a couple of metres away from the robot because beyond this range each leg begins to be seen as a single laser data point (the laser range finder we use has a range of about 4 metres).

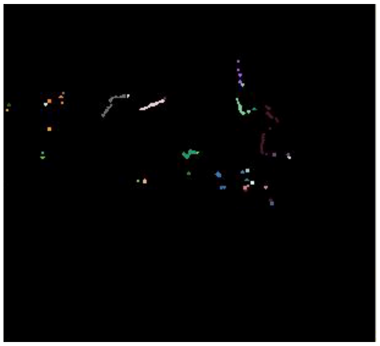

Figure 3 Segmented laser scan. Laser scanner is in centre of image. Colours distinguish segments

Figure 4 Hypothesized leg segments

For delineating the target in the image frame, simple background subtraction is used, with the assumption that both the robot and the environment (as seen by the camera) are stationary to begin with. The greyscale values of an initial frame form the background model.

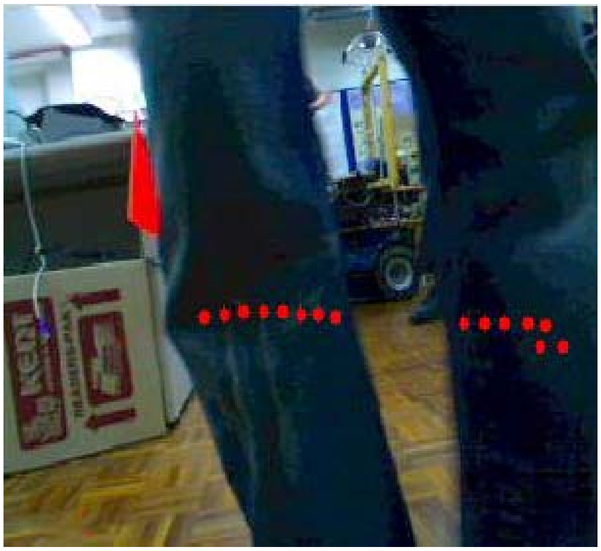

Figure 5 Laser leg segments projected onto image

Successive frames are subtracted from the initial frame, and any pixel whose intensity differs from the background model by a pre-defined threshold is detected as foreground. When a person walks into view, he/she is detected as foreground in the camera image. When within a certain distance of the robot (about 2 metres), the person is also successfully detected in the laser scan. The laser scan is projected onto the camera image, and a vertical strip is constructed as a mask within the limits of the projected leg extremities in the image (Figure 6). This strip is used as a mask into the foreground image to extract the colour model of the person's legs without background pixels. It can be seen that without the laser mask, the extraction of the legs in the image is quite noisy and includes a significant number of background pixels.

2.2 Target Colour Modelling

The colour-based leg segmentation algorithm was developed from Cai and Goshtaby's technique for modelling skin colour in chrominance space [Cai and Goshtasby, 1999]. The probability of each pixel being ground colour is determined by comparing the pixel colour in CIE Lab colour space to a colour histogram. The algorithm is summarized below:

Colour modelling:

1. The foreground image extracted by a combination of laser leg segmentation and background subtraction in the image plane is used as the colour sample.
2. Colour values in this image are converted from RGB to CIE Lab space and a histogram is computed, showing the frequency of occurence of each value in the chrominance space (a-b space).
3. The 2-dimensional histogram in a-b space is convolved with a Gaussian to obtain the "colour cloud" in chrominance space (Figure 7). The

higher intensities show the regions of the colour space with higher probabilities of being target colour.

Colour finding:

4. This “colour cloud” is then used as a look-up table to determine the probability of an input pixel from every successive input image. (Note: The input pixel needs to be first converted to CIELab space). This intensity-coded probability image is called the colour probability image.
5. The colour probability image is dilated then thresholded to get the final back-projection image.

Once the colour model is computed, the generation of the colour probability image is extremely efficient as it is just a look-up table.

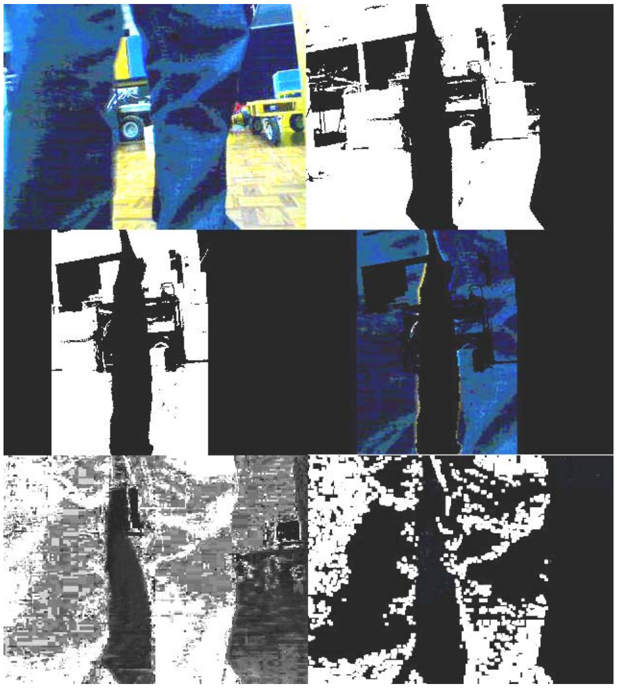

Figure 6 Steps in buildup of colour model. Left to right raster from top left: Original image, binary foreground image, binary foreground image anded with laser mask, colour samples used for modelling ,colour probability image, backprojected image

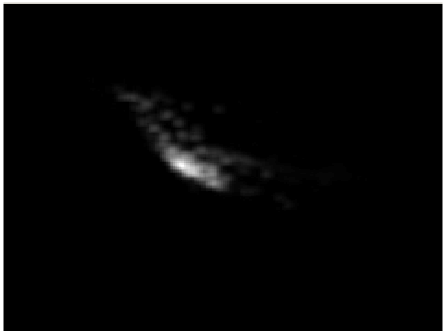

Figure 7 Multi Gaussian colour model in CIE Lab space

2.3 Target Colour Modelling

The number of “on” pixels in each column of the backprojection image is taken as one indicator of the presence of the target’s legs. This signal is plotted as a function of the columns of the image in Figure 8. The two peaks at the position of the legs illustrate that the projection of backprojected pixels onto the x-axis can be used probability density of the leg locations as a function of the x-axis. Due to the noisy nature of the signal (the maxima jumps between the two legs), a one dimensional particle filter is used to keep track of the most likely single-leg position (column) in the image. This strategy therefore works regardless of the target’s pose relative to the robot (i.e. whether 1 or 2 legs are visible).

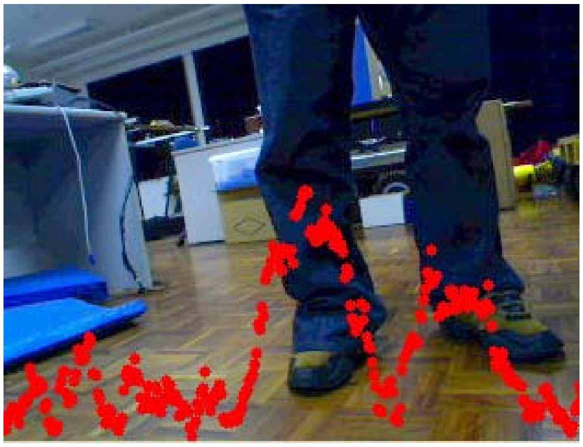

Figure 8 Plot of number of leg-coloured pixels in each column of the image

The particle filter uses a set of N = 1000 probability-weighted particles to keep track of the pdf of the leg position over time. Particles are initialized randomly over all the columns in the image. For each incoming image, the following is iterated:

Update

Each particle’s weight is evaluated based on the number of “on” pixels associated with it.

Resample

Roulette-wheel selection is used to resample 96% of the particles; particles with larger pdf having greater chance of reselection. The remaining 4% of the particles are redistributed randomly throughout state space. Random redistribution helps to prevent a total coalesce on an inferior solution.

Diffuse

Particles are perturbed from their positions by a small noise factor drawn from a normal distribution.

Extraction of Best Estimate

The best leg estimate is taken to be the mean column location of the particles (blue vertical line in Figure 9, where each particle is drawn using a red vertical line) within a window around the best particle. The best particle (yellow line) has also been considered, but it keeps jumping from one leg to the other, which is undesirable. The tracker needs to be steady if it needs to give the robot a bearing of the target for person following.

The particle filter has a complexity of O(n), where n is the number of particles. The small number of particles and the one dimension to be tracked ensure that a minimum amount of computational resources are spent in each cycle.

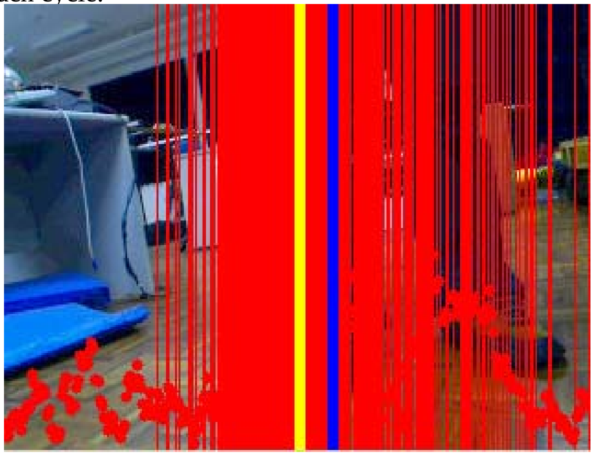

Figure 9 Particles of the one-dimensional particle filter shown as red lines. Best particle position shown in yellow. Mean particle location in a window around best particle shown in blue

Clustering

The columns are clustered based on the number of leg coloured pixels in each using a k-means clustering algorithm. The number of clusters is fixed apriori to 4. Figure 10 shows the clustering in an image with no legs (each column with a coloured point at a height indicating number of leg-coloured pixels in the column, the colour representing membership in a particular cluster). The vertical coloured bars indicate the mean height of the columns of each cluster. As soon as a leg enters the scene, the mean height of the highest cluster jumps up (Figure 11). This property is useful to detect the presence of legs in the image, and helps the a robot re-acquire a previously lost target. It can also indicate when to perform particle filtering on the image data, which may be turned off when no legs have been found for a pre-determined period of time.

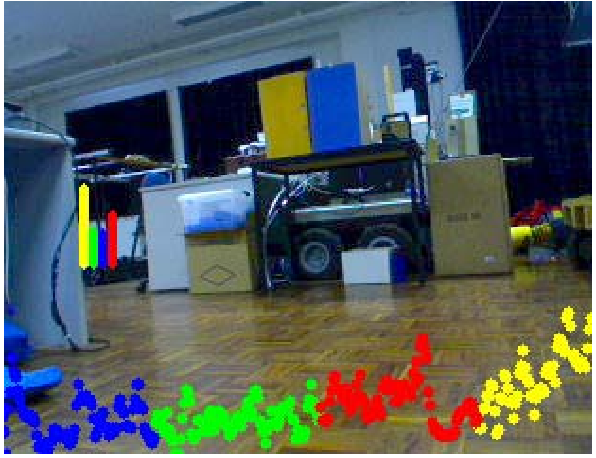

Figure 10 K-means clustering of columns based on number of leg-coloured pixels in each column. Cluster count is fixed to four. Mean values of clusters are displayed as vertical bars

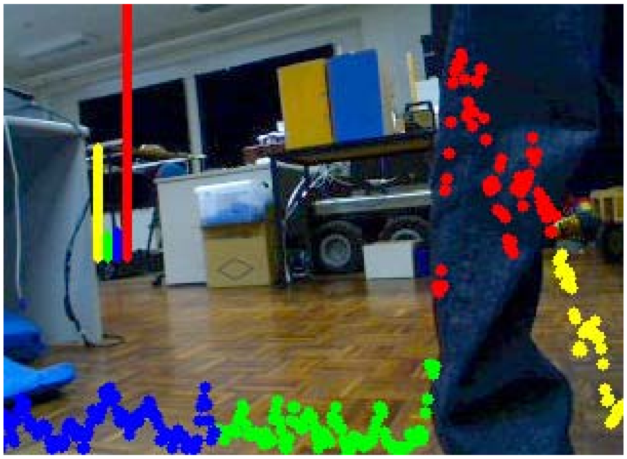

Figure 11 Mean value of cluster with highest number of leg coloured pixels shoots up as soon as a leg enters the scene. This is used to reacquire legs after lost track

2.4 Laser Association for Goal Selection

The target location, given as a column number in the image is transformed to a bearing in the laser plane by a simple linear mapping.

2.5 Navigation & Pursuit

The robot is required to navigate in realistic indoor environments thus necessitating obstacle avoidance behaviour. However, using a fixed camera for visual target tracking requires that the robot be oriented towards the target whenever possible.

Arbitration

This dichotomy of activity is resolved using a behavioural dynamics arbitration scheme based on [Althaus and Christensen, 2003] and omitting all but obstacle avoidance and goto behaviours. In our application the last known visual target bearing is continually mapped onto the robot odometric plane to become the goal for the goto function. If the visual target is within range of the laser the actual location is known and selected; but otherwise a goal is formed at the limit of laser range using only target bearing.

The result is that when dangerously close to obstacles the robot is willing to lose sight of the target temporarily before turning back towards the target's last known location. If the target moves to become obscured by intervening obstacles, the robot moves to the last location at which the target was observed thereby being most likely to successfully resume pursuit.

3 Experiment

15 trials were conducted in each of which the stationary robot encounters a hitherto unknown human "intruder"; several human actors were used, each of whom wore different clothing (however, none wore clothing that wholly obscured the form of the legs e.g. skirts). The robot was provided with no environmental information besides that collected by its onboard sensors prior to appearance of the target.

The trials were documented using screen capture and an offboard video camera moved along the robot. For the results in Figure 12, motion was recovered manually using the grid texture of the laboratory floor.

In 12 of the 15 trials the robot was successfully able to pursue the intruder for a considerable distance around a cluttered laboratory environment. In a further 5 trials the robot was physically prevented from continuing the pursuit and was thereafter successful in reacquiring the target and resuming pursuit.

It must be mentioned that although the human target actors were sympathetic to the robot by moving slowly their motion was unconstrained and undirected. Limitations on robot speed were not due to processing complexity but rather the rate of update of the scanning laser rangefinder.

Failures in 3 trials were caused by: Collision with objects invisible to the scanning laser, poor foreground selection during colour model formation and loss of target visual bearing due to sudden evasive manoeuvres.

The videos (from both an on-board screen capture system and an external camera) are available at http://users.monash.edu/~pcha25/acra06/acra06.htm.

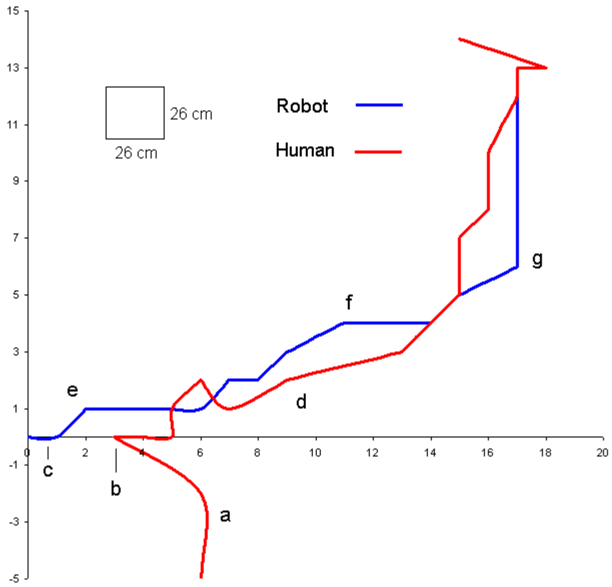

Figure 12 Example motion of Robot pursuer R and Human target T recovered from one of the trials. a) T moves into view of R and within laser range allowing R to form colour model of T. b) T spots R, and backs away. c) R advances towards T. d) T retreats quickly. e) Already R has moved away from intervening obstacles, whilst keeping T in view. f) R forced to take wide berth around desk, rather than moving directly towards T. g) Once around the obstacle, the robot moves towards T again. Tracks of T & R recovered from video of experiment. 26 cm x 26 cm unit size based on grid markings on floor. Video available at: (http://users.monash.edu/~pcha25/acra06/acra06.htm)

4 Conclusions, Future Plans

Laser and vision information was fused to detect and model the appearance of a person appearing in the field of view of a stationary robot. Target legs detected in the laser were modelled by using colour information from the camera. The legs were then tracked using a one-dimensional particle filter giving bearing information of the target with respect to the robot.

A behavioural dynamics scheme for arbitration of obstacle avoidance and target tracking activity successfully achieved reasonable pursuit performance without complex panning camera apparatus, and safe navigation in a cluttered laboratory environment without construction or provision of a detailed map. Further, the chosen solution is computationally cheap.

The current target modelling phase requires the robot to be stationary and is only done at start-up, which works well for an indoor laboratory environment with consistent lighting, but would be unsatisfactory in environments where lighting changes from place to place. This will require continuous update of the colour model. At the time of writing, depth information from stereo is being experimented with as an additional cue for the particle filter. At the moment, the target is only tracked in the visual field of the camera. Future improvements will include the ability to track the target in the laser plane when it is out of the field of view of the camera.

Perhaps the most serious drawback to the approach is that the low position of the camera makes invisible some of the most distinctive visual features of human targets – such as faces – and thus is not suitable for crowded scenarios. However, in many security and domestic applications the system might be expected to perform robustly, especially if colour models were updated at regular intervals.

Our idea is to address the problem of pursuing a single undefined humanoid target in a realistic but not busy environment, as is typical of many security applications. We desired to keep the proposed system computationally simple so that it could be deployed very cheaply on a very simple mobile platform. The choice of the background subtraction to define the appearance of the target in the 2D visual plane does not impose unecessary contraints on appearance. Since navigation is required for pursuit and intervention tasks, we allow the use of basic navigational sensors (in this case a 4-m range scanning laser) to improve the target colour model. The result is a simple visual system that is capable of target pursuit using a computationally trivial reactive navigational algorithm.

5 Acknowledgement

The API for laser scanning was sourced from Gideon Kowadlo, source code and documentation for which is available from http://users.monash.edu.au/~gkowadlo/

References

[Althaus and Christensen, 2003] P. Althaus and H. Christensen. Behaviour coordination in structured environments. *Advanced Robotics*, 17(7):657-674.

[Cai and Goshtasby, 1999] J. Cai and A. Goshtasby. Detecting human faces in colour images. *Image and Vision Computing*, 18:63-75.

[Chakravarty and Jarvis, 2006] P. Chakravarty and R. Jarvis. Panoramic vision and laser range finder fusion for multiple person tracking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006.

[Haritaoglu, *et al.*, 2006] I. Haritaoglu, D. Harwood and L.S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(8):809-830.

[Jung and Sukhatme, 2004] B. Jung and G. S. Sukhatme. Detecting moving objects using a single camera on a mobile robot in an outdoor environment. In *International Conference on Intelligent Autonomous Systems*, 2004.

[Morate, 2005] A. Morate. People detecting and tracking using laser and vision. Technical Report IR-RT-EX-0512, KTH, 2005.

[Rawlinson, *et al.*, 2004] D. Rawlinson, P. Chakravarty and R. Jarvis. Distributed visual servoing of a mobile robot for surveillance applications. In *Australasian Conference on Robotics and Automation (ACRA)*, 2004.

[Treptow, *et al.*, 2006] A. Treptow, G. Cielniak and T. Duckett. Real-time people tracking for mobile robots using thermal vision (available online at http://aass.Oru.Se/~gck/papers/at06ras.Html). *Robotics and Autonomous Systems.*

[Turk and Pentland, 1991] M. Turk and A. Pentland. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586-591, 1991.

[Zhang and Kodagoda, 2005] Z. Zhang and K. R. S. Kodagoda. Multi-sensor approach for people detection. In *International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, 2005.