Tracking People with Networks of Heterogeneous Sensors

Alex Brooks and Stefan Williams
ARC Centre of Excellence in Autnomous Systems
School of Aerospace, Mechanical, and Mechatronic Engineering
University of Sydney
NSW Australia
{a.brooks,s.williams}@acfr.usyd.edu.au

Abstract

This paper describes the theory and implementation of a system of distributed sensors which work together to identify and track moving people using different sensing modalities in real time. Algorithms for detecting people using cameras and laser scanners are presented. A Kalman Filter is used to fuse the information gathered from the various sensors. Access to information from different kinds of sensors makes each individual sensor more powerful. Results of these techniques to tracking of a person moving in a typical office environment are presented.

1 Introduction

Machines are currently very good at serving us across well-defined artificial interfaces such as keyboards and monitors, but relatively poor at interacting with us on our own terms. For intelligent machines to be integrated into our environment in a way that is natural for us they must be able to identify and keep track of us, Mimicking humans' phenomenal ability to do this has proven to be a challenging task for mobile robots.

Since networking technology has become increasingly reliable and cheap, there is no reason in principle why a robot's intelligence should be confined to a single processor or even platform. Accordingly there is a great deal of active research in the area of Distributed Sensor Networks [Qi *et al.*, 2001]. Potential advantages of tracking people with networks of sensors as opposed to a single sensing platform include better spatial coverage, robustness, survivability and modularity. Sensing accuracy can be improved by overlapping the fields of view of widely separated sensors, allowing for easier triangulation. More importantly, sensors with complimentary modalities can be chosen, allowing richer information to be extracted from the environment.

A robot with access to the information provided by a network capable of sensing the people around it will be able to interact with those people, despite carrying relatively poor sensing and processing power on board.

To demonstrate some of these ideas, a system that tracks people in real-time by fusing information from laser and camera sensors has been developed. These sensors are complimentary in that information gained from one sensing modality is used to validate observations from the other. In addition, this selection of sensors allows information about the state of the environment to be deduced that would unavailable to either sensor working alone. For example, accurate range and bearing measurements from a laser are used to bound the likely position of a person moving in the scene. This gives no information about the height of the person or their appearance. Height estimates are then generated using observations from a monocular camera. By fusing these two observations the height of the person can be resolved and they can be tracked in real time. Observations are currently fused in a central extended Kalman Filter.

The rest of the paper is organised as follows: Section 2 describes previous work in the field of people-detection and tracking. Sections 3 and 4 describe techniques used to interpret the data from lasers and camera respectively, and Section 5 details the techniques used to fuse this information. Finally, Section 6 provides some experimental results.

2 Previous Work

There is a great deal of literature on object-tracking in general and people-tracking in particular using various sensors, the most popular subject being people-tracking in video sequences. Many approaches use quite complex models of their subjects, resulting in a need for large amounts of processing power. This thirst for processing power renders these approaches unsuitable in the case where the outputs from multiple sensors must be processed with small, distributed processors. Instead this paper focuses solely on fast approaches such as colour-based tracking and approaches based on background-subtraction.

Human-finding based on skin-colour [Vezhnevets, 2002] is fast, however it has fundamental problems with people wearing shorts or short-sleeved shirts, and with skin-coloured clothing or objects such as wood. Fieguth and Terzopoulos [Fieguth and Terzopoulos, 1997] improves robustness by introducing extra constraints such as the spatial relationships between coloured regions, however this constrains the set of objects that can be tracked to those that are reasonably rigid.

Wren et. al. [Wren *et al.*, 1997] use background and foreground classes to distinguish between foreground blobs and a fixed background in real-time, based on spatial and colour properties. A single fixed camera is used, and tracking is limited to a single target. Unlike our work, PFinder tracks particular body parts for use in gesture recognition.

Comaniciu et. al. [Comaniciu *et al.*, 2003] present a real-time colour-based system for tracking objects in an image stream based on a smooth similarity measure which defines a basin of attraction for a local maxima corresponding to the object's location. The system doesn't require a static background, but does require some form of initialization, currently performed by a human. It also requires that the tracker begin its search for the target within the basin of attraction. This is trivial with a high enough frame rate or slow enough motion, but requires some sort of prediction otherwise.

Besides colour-based tracking, the only other image-processing technique that has been shown to be fast, reliable and effective has been segmenting motion based on background subtraction [Haritaoglu *et al.*, 1998] [Jabri *et al.*, 2000]. Unfortunately these methods require that the background be reasonably static and that the camera be fixed.

In order to determine people's locations in the world, location in a 2D image is insufficient. We consider stereo vision to be too computationally expensive and unreliable for our purposes and the techniques do not easily scale to multiple, distributed and heterogeneous sensor systems. Other approaches include sensing range directly with non-vision sensors, and triangulating using multiple sensors.

Scanning laser range-finders have been used on their own to track people [Fod *et al.*, 2002] [Fuerstenberg *et al.*, 2002] [Prassler *et al.*, 1999]. In contrast to cameras, lasers give highly accurate depth information and require very little processing since they return so little data. Further, they are insensitive to noise from sources such as ambient lighting. However they suffer from only being able to sense in a plane, and extract only very primitive intensity information and no colour information.

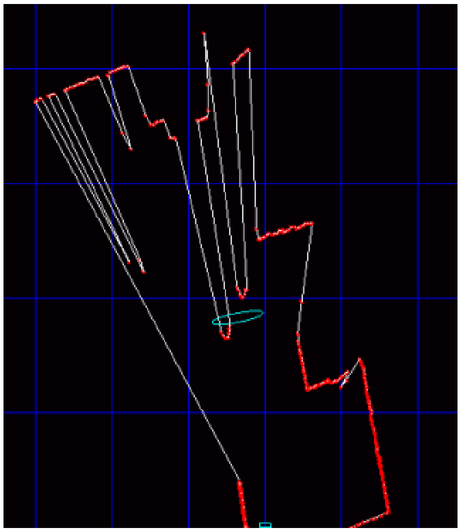


Figure 1: Raw laser scan in which two legs can be distinguished. The estimate of the person's location is given by the cyan ellipse.

Dockstader et. al. describe a system that recovers the locations of tracked objects in world coordinates by triangulating from several cameras [Dockstader and Tekalp, 2001]. A Bayesian Belief Network is used to generate 3D observations which are then fused by a central Kalman Filter. Stillman et. al. use a combination of static cameras for general person detection and PTZ cameras to follow and identify faces [Stillman *et al.*, 1998].

The task of locating and tracking people with heterogeneous sensors has received some attention in the context of 'Intelligent Spaces' or 'Aware Homes', specifically with cameras and microphone arrays [Trivedi *et al.*, 2000] [Stillman and Essa, 2001].

3 Sensing With Lasers

A SICK laser rangefinder, mounted such that the sensing plane is horizontal and about midway between the knees and ankles on an adult, is used to take observations of people moving in an area. This configuration was chosen as a pair of legs represent a fairly distinct feature when viewed in a plane whereas other parts of the body are not as easy to discriminate. A leg, as sensed by a laser scanner, is modelled as a semi-circle with radius in a certain range, separated from its background by a threshold distance determined experimentally. A human is defined as a pair of legs within a certain distance of each other. In the lab environment in which this system is operating this approach to identifying people produced very few false positives.

Once a scan has been processed to identify likely legs, a range bearing observation to a point midway between the two legs is generated. Figure 1 shows the detection of a pair of legs in a laser scan. The approach is computationally extremely light. The 2D shape of the person needn't be modelled explicitly since leg shape is fairly invariant. Another advantage of this approach is that it is suitable for use on a mobile robot without modification whereas methods that rely on background subtraction need to estimate sensor motion [Prassler *et al.*, 1999]. This scheme does have the disadvantage that it fails to detect humans when one leg occludes the other. Temporary occlusions are handled by the Kalman Filter, especially when complemented by observations from other sensors in the network.

4 Sensing With Cameras

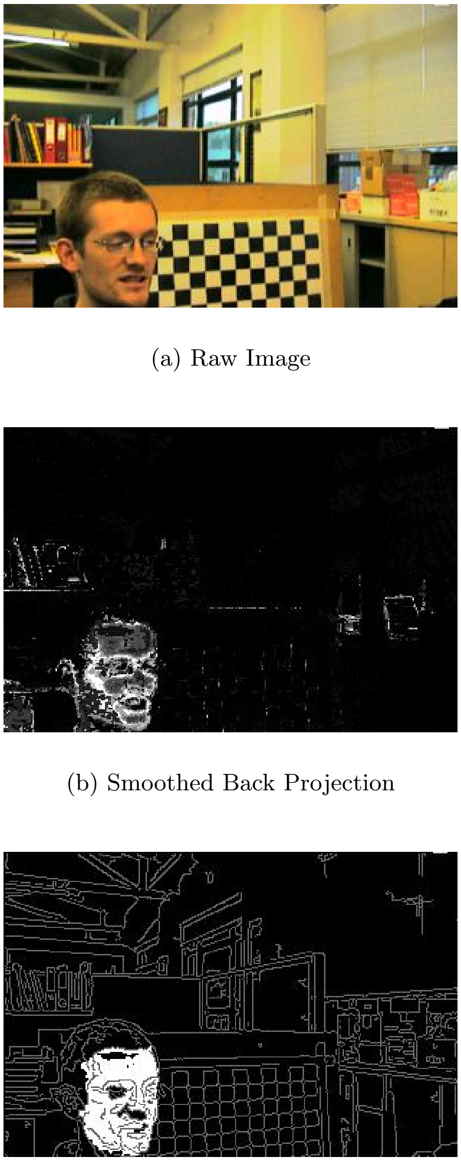
Humans are identified in the camera images by looking for moving objects with skin-coloured blobs (corresponding to heads) on top. Moving blobs are segmented using background subtraction and a potential target is represented as the blob's bounding rectangle. These rectangles are validated by requiring that they have some proportion of skin-coloured pixels in their top fifth.

The background subtraction method adopted for this work is based on the method proposed by Jabri et. al. [Jabri *et al.*, 2000]. Incoming frames are split into colour and edge planes. For each plane, the mean and standard deviation of each pixel is calculated recursively as a weighted average of the mean and standard deviations over the previous few frames, using an exponential forgetting factor. By monitoring standard deviations, persistent movement from things like flickering computer monitors can be ignored to some degree. Slowly updating the means also allows gradual changes in lighting conditions to be accounted for.

This approach is especially suited to limited processing power because it can scale by simply adjusting the number of planes into which images are split. For example colour can be analysed in 3 planes (R, G, and B) or in grayscale. Edges can be analysed in one dimension using a Laplacian operator, in two using horizontal and vertical Sobel operators, or can be ignored if required. Experimentation shows that performance scales accordingly which is an important consideration as our goal is to deploy these sensors on nodes with limited computational power.

Skin colour is segmented using the normalized-RGB colour space because it has been shown that human skin colours cluster in a small region in this colour space, characterized by a multivariate gaussian distribution [Yang *et al.*, 1998]. This result does not depend heavily on the human's race or amount of tanning, but is affected more seriously by the ambient lighting conditions.

A model for skin colour is generated off-line from an example image, taken with the camera in the lighting conditions in which it is to be used. Skin patches are


(a) Raw Image

(b) Smoothed Back Projection

(c) Segmented Image

Figure 2: Skin Segmentation Procedure. The back projection is produced using a histogram of skin colours. The segmented image is produced with hysteresis thresholding. The blobs are limited by edges, indicated by the gray lines.

manually segmented in the image and a colour histogram is built from all the segmented pixels.

When the system is put on-line, the amount to which an incoming pixel resembles skin is defined as the value of the pixel's corresponding bin in the normalised-RGB colour histogram. This procedure, called Back Projection, produces an image which is a map of likely skin regions. This image is smoothed and then hysteresis thresholding is applied, bounded by edges obtained using a canny operator on a greyscale version of the current frame. Figure 2 shows an example of this skin-segmentation procedure.

5 Fusing Heterogeneous Sensor Data

Each sensor independently produces observations which are then fused into the global state using an Extended Kalman Filter. Targets are modelled with three state variables: x-y position in cartesian coordinates, and height.

Velocity isn't modelled for various reasons. To model a human as a constant-velocity point target is fundamentally difficult because purposive human behaviour doesn't maintain anything like a constant velocity. Furthermore, modelling velocity is problematic due to the rate at which observations are made. The laser produces a scan every 0.2 seconds and the camera takes a similar amount of time to process an image due to limitations in computing power. Since these periods are of a similar order of magnitude to the time taken for a human to accelerate from zero to maximum velocity, process noise needs to be large enough to account for large changes in velocity between observations. This large velocity noise means that after a second or two with no observations (due to occlusion for instance) the Filter has almost no idea where the target could be. The three-sigma covariance ellipse grows to cover a huge area. This problem could possibly be solved by using multiple cameras such that their observation rates add. Instead of modelling veolocity, targets are simply modelled as points with constant position plus gaussian noise. Refinement of the vision algorithms to include an estimate of the heading of the person may allow us to produce more accurate motion models but the current approach has yielded satisfactory results during initial testing.

Both laser and camera observations are required for the robust operation of the system. Camera observations alone will produce a relationship between height and range, but their true value cannot be resolved. Absolute range observations, as provided by the laser, are required to disambiguate the scale of the scene.

Height is explicitly modelled to allow cameras to observe depth if the laser loses track of the target. The mounting height and angle of cameras are currently measured prior to deploying the system. Using an estimate of camera height, observations of the angle of elevation to the top of a target can then be transformed into observations of range using simple geometry under the assumption of level ground. In the absence of laser observations, some form of range observation is required because bearings-only observations will tend to draw estimates of targets' positions toward a sensor. Consider repeated bearings-only observations of an erratically moving target. Underestimating the target's range results in estimates of smaller sensor and process noise, and therefore a more probable hypothesis. Range estimates obtained from height, while not exceptionally accurate, tend to counteract this effect. Viewing the target from multiple vantage points also aids in this respect.

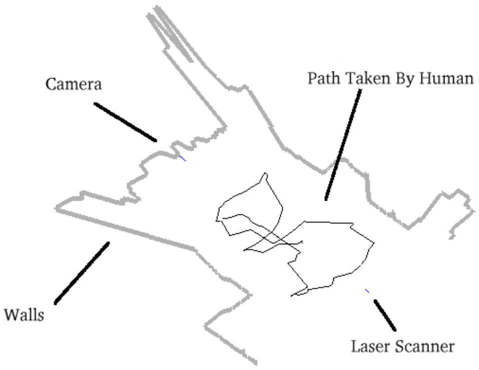


Figure 3: A view of the test area from above, showing the sensors, walls and path of the human as estimated by the tracker.

The height model is also used for distinguishing between targets. The standard three-sigma innovation gate is used for validating observations. This allows the tracker to ignore observations which would result in an unrealistic estimate of height.

6 Experimental Results

The system described above was implemented for a camera and a laser scanner located on opposite sides of an area in a lab environment, both sharing a view of the same area. A view of the area from above, showing the walls, sensors and path of the human as estimated by the tracker is shown in Figure 3. While the sensors are distributed, the processing is currently performed on a single computer. When a human is first identified the Kalman Filter is initialised with the unrealistic height of 1.0 metres. Uncertainties in both position and height are initialised to be very large, and drop off dramatically after a few observations.

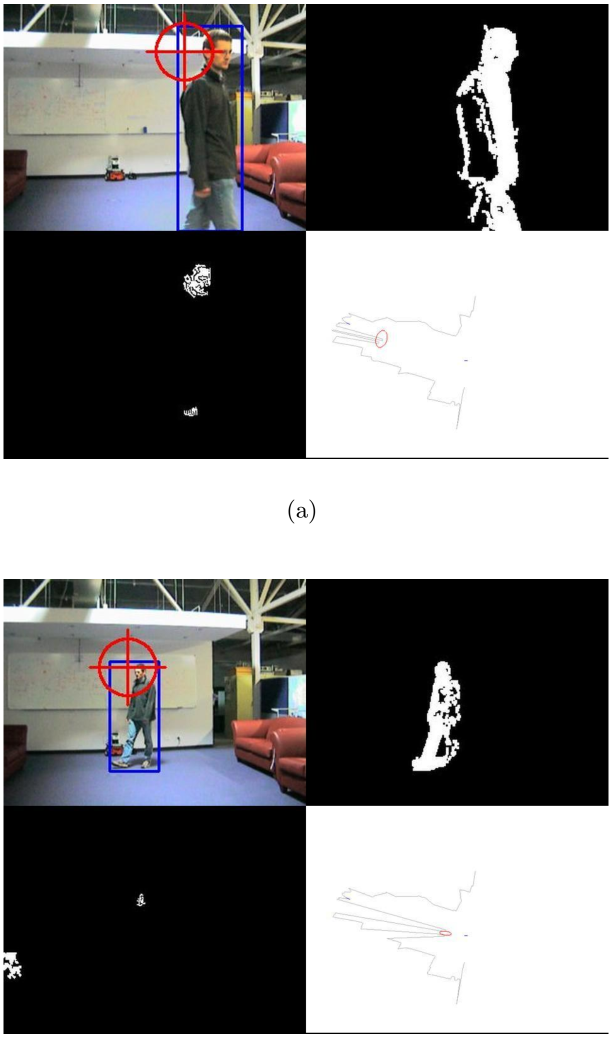
Some example frames from a typical run are shown in

Figure 4. Each frame has four components: the upper-left component shows the raw image from the camera, with the direction to the top of the human indicated by red crosshairs. The crosshairs show the current prediction of the Kalman Filter, so they are always present regardless of whether or not a human was detected in the current frame. Instantaneous observations in each frame are shown by drawing a blue bounding box around the human. The upper-right component shows the movement that was segmented from the image and the lower-left frame shows segmented skin-colour. Finally, the lower-right component shows the output from the laser scanner, seen from above, with the laser scanner located to the right of the image. A red ellipse is drawn around the three-sigma bounds of the human's location.

Evaluating the performance of the tracker is difficult because ground truth of the person's motion is not available for comparison with the estimate generated by the filter, and representing that truth with a point target is an approximation. Instead, correct operation of the tracker is verified by watching the laser scan. The location of the human is very obvious from this image, so it is trivial to see that tracking is approximately correct across the entire sequence (see Video 1). Due to the fact that velocity was not modelled, position estimates lag slightly behind moving targets.

At the end of the sequence the target's height was estimated to be 1.87 metres, compared with the value of 1.84 metres which was measured with a tape measure. Figure 5 shows how this height estimate evolved over the 60-second trial. For the first few seconds, before the first observation, the height remains extremely uncertain but converges fairly rapidly. Figure 6 shows the behaviour of the height estimate, including 3-sigma uncertainty bounds, starting from the time of the first camera observation.

To show that this information can be used by the camera a second trial was then performed in which the human was tracked with the camera only, using the height information that had been gathered in the previous trial. The laser scan is still shown to allow assessment of the tracker, however the scan is unavailable to the tracker. This experiment showed that the human can still be tracked, but with less accuracy (see Video 2). This video shows that when the human is first detected, the bearing is known relatively accurately but the range is extremely uncertain. This is because observations of range based on height are fairly inaccurate. After a few more observations the tracker resolves the range fairly well, but not as accurately as the camera and laser working together. The only point at which the system loses track of the human is when he is in a position such that the skin colour of the face can't be detected. Sensing with multiple cameras from different vantage points would reduce



(a)

(b)

Figure 4: Two sample images from a tracking sequence. The crosshairs show the current filter estimate, and the blue bounding box shows the current observation. Segmented skin colour is in the lower-left component, and motion is shown in the upper-right. The laser scan is shown in the lower-right, with the target shown in red.

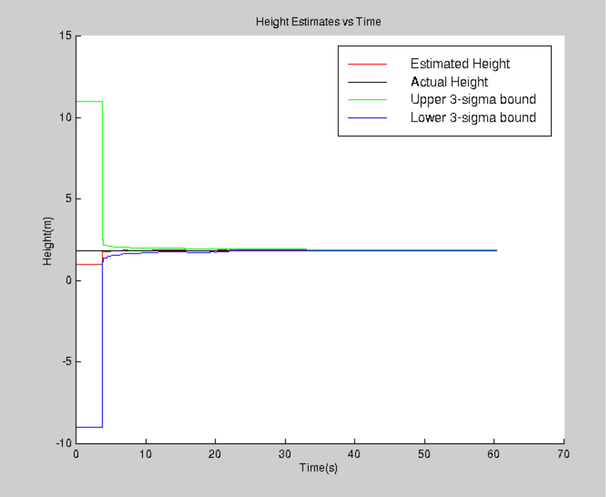


Figure 5: The evolution of the estimate of the target's height over time.

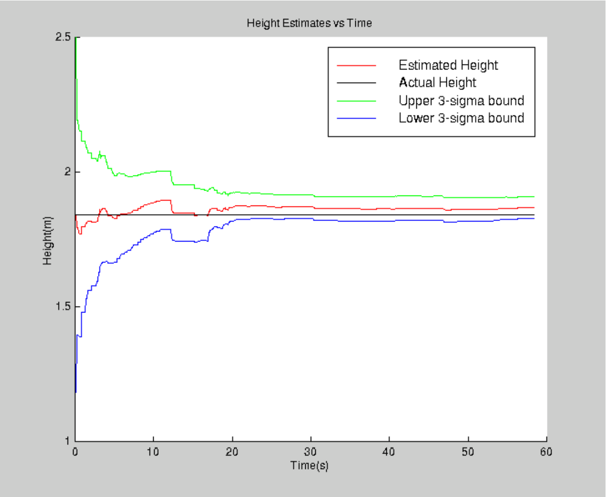


Figure 6: The evolution of the estimate of the target's height over time, starting from the first camera observation.

the likelihood of errors such as this.

7 Conclusion and Future Work

This paper has presented a system in which information from imperfect, heterogeneous, physically distributed sensors is fused to track people walking in a lab environment. It has been shown that observations from individual sensors can be improved based on the information gained from other sensors.

Ongoing work is aimed at distributing the processing across multiple processors and removing the need for a central fusion unit by using Distributed Data Fusion techniques [Durrant-Whyte and Stevens, 2001].

8 Acknowledgements

This work is supported in part by the ARC Centre of Excellence programme, funded by the Australian Research Council (ARC) and the New South Wales State Government.

References

[Comaniciu *et al.*, 2003] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, May 2003.

[Dockstader and Tekalp, 2001] Dockstader and Tekalp. Multiple camera fusion for multi-object tracking. In *IEEE Workshop on Multi-Object Tracking*, pages 95–102, 2001.

[Durrant-Whyte and Stevens, 2001] H. Durrant-Whyte and M. Stevens. Data fusion in decentralized sensing networks. In *4th International Conference on Information Fusion*, pages 302–307, 2001.

[Fieguth and Terzopoulos, 1997] P. Fieguth and D. Terzopoulos. Color-based tracking of heads and other mobile objects at video frame rates. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 21–27, 1997.

[Fod *et al.*, 2002] Fod, Howard, and Mataric. Laser-based people tracking. In *Proc. IEEE Intl. Conf. on Robotics and Automation*, pages 3024–3029, 2002.

[Fuerstenberg *et al.*, 2002] K. Fuerstenberg, K. Dietmayer, and V. Willhoeft. Pedestrian recognition in urban traffic using a vehicle-based multilayer laserscanner. In *Proceedings of IV 2002, IEEE Intelligent Vehicles Symposium*, volume 1, pages 31–35, 2002.

[Haritaoglu *et al.*, 1998] I. Haritaoglu, D. Harwood, and Larry Davis. W4: Who? when? where? what? a real time system for detecting and tracking people. *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 222–227, 1998.

[Jabri *et al.*, 2000] Jabri, Duric, Wechsler, and Rosenfeld. Detection and location of people in video images using adaptive fusion of color and edge information. *ICPR*, 4:4627–4631, September 2000.

[Prassler *et al.*, 1999] E. Prassler, J. Scholz, and A. Elfes. Tracking people in a railway station during rush-hour. In *CVS99*, pages 162–179, 1999.

[Qi *et al.*, 2001] Hairong Qi, S. Sitharama Iyengar, and Krishnendu Chakrabarty. Distributed sensor networks - a review of recent research. *Journal of the Franklin Institute*, pages 655–668, 2001.

[Stillman and Essa, 2001] S. Stillman and I. Essa. Towards reliable multimodal sensing in aware environments. In *Perceptual User Interfaces (PUI 2001) Workshop*, 2001.

[Stillman *et al.*, 1998] S. Stillman, R. Tanawongsuwan, and I. Essa. A system for tracking and recognizing multiple people with multiple cameras. Technical report, Georgia Institute of Technology, 1998.

[Trivedi *et al.*, 2000] M. Trivedi, I. Mikic, and S. Bhonsle. Active camera networks and semantic event databases for intelligent environments. *IEEE Workshop on Human Modeling, Analysis and Synthesis*, June 2000.

[Vezhnevets, 2002] V. Vezhnevets. Method for localization of human faces in color-based face detectors and trackers. In *The Third International Conference on Digital Information Processing And Control In Extreme Situations*, pages 51–56, 2002.

[Wren *et al.*, 1997] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[Yang *et al.*, 1998] Jie Yang, Weier Lu, and Alex Waibel. Skin-color modeling and adaptation. In *ACCV (2)*, pages 687–694, 1998.