Introduction:

Mobile robots intended to interact with people in indoor office environments, such as the Stanford AI Robot, need to be able to detect and locate people in their surroundings. Unfortunately, the task of detecting people with vision methods is very difficult due to the wide variety of possible poses that people can take on as well as variability in body shape and haberdashery. We worked on detecting people with 3D range data and passive infrared data as a robust alternative. A side benefit of this approach is that we are capable of accurate 3D localization of people in addition to simple detection. This localization enables interactions such as fetching an object and bringing it to the person.

Related Work

There is a very large body of past work in person detection, yet none of it has been very successful. It is a very active area of research, because of the many potential applications, which include robotics, surveillance and pornography filtering.

Much research has explored the possibility of person detection via vision methods. Mohan, Papageorgio, and Poggio claim 99% recall with less than 0.1% false detection rate, using an SVM with HAAR wavelets for features. Tunzel, Porikli, and Meer claim less than 1% miss rate with less than 1 false positive per 100 frames, using covariance matrices of image brightness and derivatives of brightness as features for a custom machine learning algorithm that exploits the fact that covariance matrices lie in a manifold rather than a complete vector space. Both of these examples are somewhat misleading, however, because both place significant constraints on the pose of the person.

Other vision methods include using background subtraction to detect motion and assuming that any large, moving object is a person (Elgammal). These methods are acceptable for surveillance applications but not for mobile robots, because the background does not remain constant if the camera is moving. Other methods get very good accuracy but rely on the person's face being visible, such as Patil, Rybski, Kanade, and Veloso's tracking system.

The use of range data to detect people has been very limited so far. Arras, Mozos, and Burgard search for the shape of knees in 2D range data from a SICK laser, and achieve a maximal detection rate of 90%. Xu and Fujimura use the same 3D data that we use, but only search for ellipses of hand-coded size. Their system does not use any machine learning and they do not publish any quantified results.

Much work has also been done with infrared person detection. In general, infrared person detection suffers badly from false positives. Even the best infrared detection systems, such as the one due to Fang, Yamada, and Ninomiga, can get false detection rates that exceed true positive rates. Even with low detection rates, their false positive rate can be as high as 100% in some environmental conditions.

To our knowledge, the combination of infrared and depth has only been explored previously by Bertozzi, who uses stereo vision in both infrared and the visual spectrum ("tetravision") to get depth estimates. The maximal detection rate of his system is in the 80-90% range.

Apparatus:

The 3D range data is provided by a SwissRanger 3100 3D camera, which returns entire point clouds at video frame rates through a time-of-flight method. It also provides a grayscale image of near-IR reflectivity of the scene at a resolution of 176x144 pixels. The IR data is provided by a FLIR thermal infrared camera at a resolution of 320x240. Unfortunately, our model of infrared camera does not produce temperature calibrated output—that is, it does not return absolute temperatures but rather relative temperatures. This means that while brightness in the image is correlated with heat, we can't use any absolute thresholding features in our model (unfortunate given that healthy people fall in a very narrow temperature range).

We built a mount for the cameras that immobilizes them at a width of approximately 10 centimeters apart and facing in the same direction.

Backprojection of the IR image into the point-cloud requires accurate calibration of the two cameras. While this was not a primary focus of our research it turned out to be a more difficult problem than anticipated. Two-camera calibration techniques for stereo vision are very mature, but coming up with corresponding points quickly and accurately in such different sensor modalities turned out to be troublesome.

Our 3D camera conveniently provides a near-IR reflectivity image in addition to the point cloud. However, designing a calibration object that has features visible in both near-IR (essentially the same as a visible spectrum grayscale image) and thermal IR would have required an active source of heat and significant investment of time in designing and fabricating the object. As a compromise we used a standard calibration checkerboard and shone a strong light on it. The ink selectively absorbed much more heat and the checkerboard became visible in thermal IR. Unfortunately the heat dissipated quickly, and the resulting images were rather fuzzy.

The calibration we obtained was usable, but a better calibration would likely improve our results. Our current calibration often projects the infrared values for people onto the points of a wall behind the person in the 3D point cloud, so the correlation between infrared brightness and personhood is weakened.

Approach:

We began the project with an existing person detection system that Ian created as part of the CURIS program. It works only with 3D range data, not with infrared data, and achieved a detection rate of about 85% for a recall rate of about 85% on a relatively easy data set. The detector works in three stages:

1. 3D point cloud is filtered to remove spurious points that integration in the 3D camera inadvertently introduces at depth discontinuities. We consider a point to be an artifact if the variance of the z coordinate of the 3x3 pixel neighborhood is greater than a hand-picked threshold. These points are thrown out, as they do not generally correspond to physical objects.

2. Each horizontal scan-line and each vertical scan-line is split at the depth discontinuities discovered in the first step. This generates a collection of object cross-sections for our classifier to work on. The benefit of the segmentation into cross-sections is that it produces a small, tractable number of regions to classify and that it prevents artifact points at the edges of regions from affecting our shape features (otherwise artifacts would actually be the most salient features of the data).

3. A variety of features are then computed for these segments, and standard classification approaches are applied to the problem of determining if a given segment is a segment of a person. Several variations on the features and classifiers are described below. We train separate classifiers for different elevations (absolute y coordinate) and so as an ensemble the classifiers are capable of learning different distributions of people in different locations (people are rarely

found on the ceiling for example, so lumpy shapes up high are much more likely to be lights). Since we are targeting a mobile robotics application, we are considering data taken from a uniform height.

4. The probabilities returned from the horizontal and vertical scan-line are then used to generate features for a second layer of classification. This stage looks at the distribution of probability values of points that fall into a person-sized rectangular prism to provide a final result. This step uses the same classification algorithm as the second step, and the result is a set of probabilities for a variety of rectangular prisms in the scene containing people. After thresholding and suppression of similar results we obtain the final detections of people.

We maintained this same classification pipeline for this project, and focused on augmenting the 3D features with IR features to improve performance.

Classifiers:

For the two classification steps we initially used a Bayesian logistic classifier which had the advantage of training and classifying very quickly. We found it useful for feature selection because the regularization term drives the weights of less informative features to zero. We initially used our own implementation of logistic regression but then changed the training algorithm to use an LBFGS library for compatibility with our research group's codebase (Jorge Nocedal).

We also experimented with an SVM using the radial basis function as a kernel. For this we used the LIBSVM implementation (Chih-Chung Chang and Chih-Jen Lin).

Features:

Our 3D features for each cross section include:

- The raw depth of 10 points along the cross section. This is normalized by subtracting the mean depth of the cross section (so that shapes up close produce the same result as shapes far away). We down or upsample each cross section with a rect kernel in true 3D space (rather than pixel space) so that the same shape always produces the same collection of 10 sampled values regardless of its location.
- The width of a cross section
- Coefficients and intercept terms of 1st and 2nd degree polynomials fit to the sampled depth values
- The angle between two lines fit to opposite halves of the cross section

Earlier experiments during CURIS found that with these features, many other features become redundant (their weights are driven to zero by regularization, or they do not improve classifier performance). These include fitting ellipses to the data or including features for the error between the true values and the shapes fit to them.

Our infrared features include:

- The mean infrared brightness of a segment
- The mean normalized infrared brightness of a segment, where the brightness at each pixel is normalized by subtracting the mean brightness of the image and dividing by the variance of the image. This is to compensate for drift in the camera's output over time.

•A 5 bucket histogram of the brightness of the image points in the segment. This is to compensate for cases where the calibration projects only half of the image of the person onto the point cloud for the person. The histogram can discriminate between a region of uniformly medium brightness and a region that is half dark and half bright.

Additionally, when using the logistic we also include some of the feature values squared, so that the classifier can learn a two-sided boundary on that feature. We limit this only to features that we find it to be useful for, such as width and brightness, because adding too many features begins to cause overfitting rapidly.

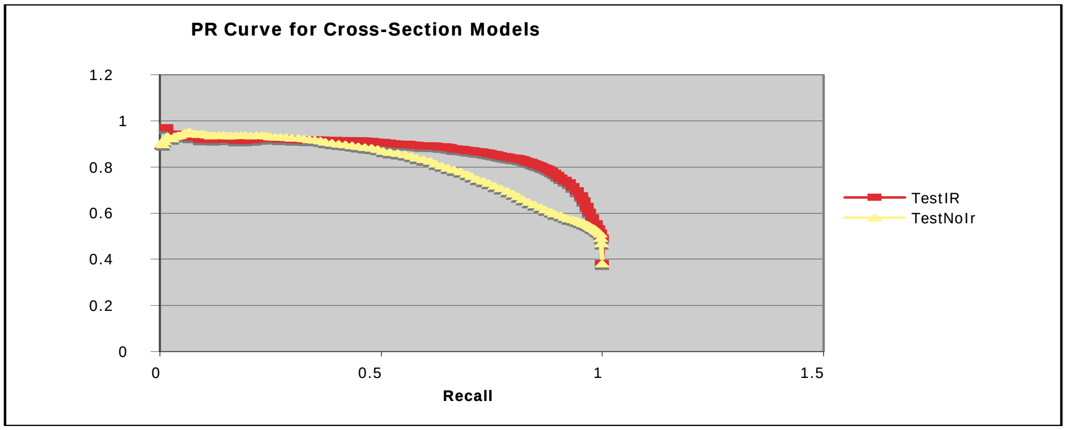
Data

We captured over 300 image pairs and labeled 200 of them. The new images are considerably more difficult than the ones Ian used for the CURIS project. We captured scenes with a greater variety of backgrounds than had been used before, and were sure to include several scenes that would disrupt the 3D segmentation algorithm, such as images of people holding camera tripods that bisect their body. We also had a much wider range of heights in this data, and had many images of mainly occluded people, such as people sitting down in a chair with their back to the cameras and only their head and shoulders visible.

When experimenting with only the cross-section classifiers we used 140 for training and 60 for testing. When experimenting with both layers, we used 80 to train the cross-section layer, 80 to train the second layer, and 40 for testing.

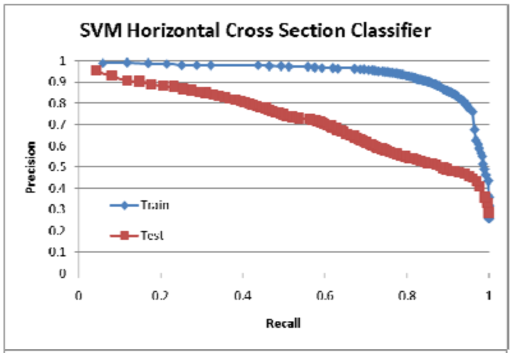
Results

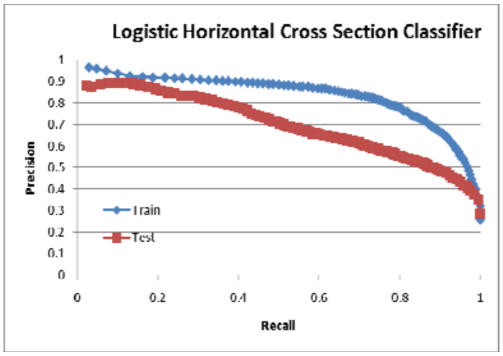
Our most important result was to demonstrate an improvement in our detection accuracy by incorporating infrared features. The following PR curve demonstrates a significant improvement at most areas along the curve:

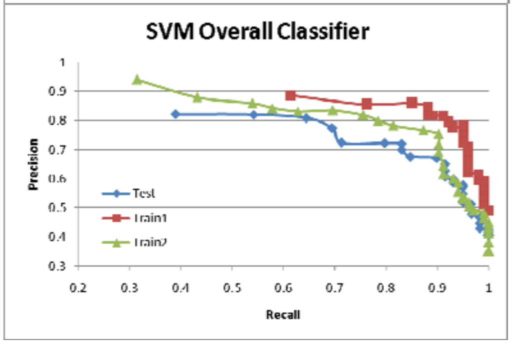


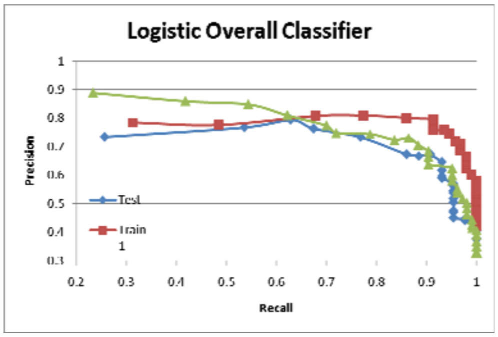
We think that we could see a much stronger improvement with better ways of normalizing the infrared values and better calibration.

Other experiments confirmed that our logistic models work nearly as well as an SVM using the RBF kernel. This is good to know because it means that we can continue to depend on the logistic for development and feature evaluation.









One thing that is troubling about these results is that the performance of the second layer of classifiers is worse than the first. In other words, we would be better off using our local probability map to produce our final decision on where people are located than we are using our probability map that was meant to evaluate the probability of a rectangular prism region containing a person. This indicates that the features we extract from the prism discard too much information.

Conclusions and Future Work:

The use of both IR and 3D range data for person detection results in reliable detection of people of very different sizes and in difficult poses. There is significant room for improvement using our existing approach, particularly in establishing the correspondence between infrared images and 3D point clouds, and in the second classification stage. In preparation for a paper submission to RSS or AAAI, we plan to work on using gradient ascent on correlation between edge maps to automatically learn a better calibration, and to conduct experiments to determine the best way of normalizing infrared values. We expect that this will yield a very fast and reliable means of detecting people in indoor environments.

Acknowledgements

We are part of the STAIR vision team. All of our work uses the STAIR vision library. In particular, much of our code, such that used to working with camera extrinsics and intrinsics, was written by Stephen Gould.