Robust and Efficient People Detection with 3-D Range Data using Shape Matching

Daniel Hordern Nathan Kirchner
University of Technology, Sydney
{Daniel.Hordern, Nathan.Kirchner}@eng.uts.edu.au

Abstract

Information about the location of a person is a necessity for Human-Robot Interaction (HRI) as it enables the robot to make human aware decisions and facilitates the extraction of further useful information; such as low-level gestures and gaze. This paper presents a robust method for person detection with 3-D range data using shape matching. Projections of the 3-D data onto 2-D planes are exploited to effectively and efficiently represent the data for scene segmentation and shape extraction. Fourier descriptors (FD) are used to describe the shapes and are subsequently classified with a Support Vector Machine (SVM). A database of 25 people was collected and used to test this approach. The results show that the computationally efficient shape features can be used to robustly detect the location of people.

1 Introduction

The fundamental capabilities for robots to perceive the world and interact with it are maturing. These capabilities include Simultaneous Localisation and Mapping (SLAM), path planning, obstacle avoidance and exploration, which are the result of significant research in the robotics field over several decades. It is now increasingly common for robots to employ these capabilities in a wide variety of applications. The maturation and robustness of these attributes mean that robots are becoming more feasible and accessible to the general society. However, this presents new challenges for the robotics field; enabling robots to naturally and cohesively exist in the same environment as people. It is likely that the ability to interact naturally and intuitively with humans will assist with the adoption and integration of robots into society. Therefore, it is important to provide these capabilities for robots which need to interact with people.

When people interact, low-level body language, proximity, gaze, speech and other such cues provide valuable information about an interaction. This information can be used for high level understanding about an interaction, such as identifying: if someone wants to interact them, if someone wants their attention, who is likely participating in the current interaction, the speaker and the addressee. If a robot also had the capacity to make these interpretations, it would assist in resolving ambiguities, thus facilitating natural and intuitive interaction. However, the low-level information needs to be obtained to do this. It is therefore evident that there is a need to extract these human cues.

The first step towards this is robustly locating people in the environment. The location of a person is necessary in extracting important information, such as proximity, pose and gestures. The variable conditions and environments in which HRI may occur, such as the home, provide challenges for determining the location of a person, such as visual complexity and lighting variations. Changes in illumination can be caused by varying artificial and natural light and the unique lighting arrangements for different rooms. The physical environment in which people occupy contains many variable objects, such as unique types of furniture, and is structurally dynamic; people interact with their environment, e.g. move furniture around. Therefore, a method that is robust to these factors is required.

Significant work in the area of person detection has been performed in the vision field. Face detection is a common method in this domain to detect the presence of a person, [Viola and Jones, 2004; yu Lin, 2005; Waring and Liu, 2005; Rowley *et al.*, 1998; Buciu *et al.*, 2001]. Although these methods have been shown to positively detect faces, in visually complex and feature rich environments they are likely to provide false positives. A widely used and well known method for detecting people is the Haar feature based face detector, [Viola and Jones, 2004], Figure 1 shows some false positives from this method. The face detection approach also limits the ability to extract further information, such as low-level body language, as it does not provide information

Figure 1: False positives of the Haar feature based face detection.

about the rest of the person or facilitate segmentation.

Background subtraction techniques are commonly utilised in people detection and tracking tasks in surveillance applications. Although background subtraction methods can be robust to changes in illumination, [Cezar Silveira Jacques *et al.*, 2006], the illumination changes are relatively slow and smooth processes, such as changing daylight. Furthermore, background subtraction techniques only deal with the segmentation of the scene into background and foreground. Therefore these foreground objects need to be classified with a technique, such as Histograms of Oriented Gradients (HOG) [Zeng *et al.*, 2008]. The static sensing approach required for background subtraction makes it impractical for use with our dynamic robotic system.

In [Dalal and Triggs, 2005], the HOG based people detection algorithm is presented. The HOG descriptors describe the local distribution of intensity gradients in an image. The HOG features were combined with a linear SVM to detect people. An example application of HOGs is presented in [Bertozzi *et al.*, 2007] where it used for pedestrian detection. In [Li *et al.*, 2009] the HOG features are combined with a Viola-Jones type classifier to detect people using omega-like features which represent the head and shoulder region of people. The good results obtained in detecting people with this region supports its use as a feature for our approach.

An issue with many vision techniques is with the sensor itself. Changes in illumination have degrading affects on many vision applications. As has been described, natural HRI will occur in unconstrained environments, where illumination changes are likely to occur in different areas. Also, the 2-D characteristic of the camera data does not directly allow for the persons location in the environment to be fully known. This also limits the capacity to extract subsequent information; such as pose and gestures.

Whilst many of the aforementioned techniques have proven to be successful in specific applications, they are not ideal solutions for our requirements. Also, the need for 3-D information to facilitate with extracting and identifying additional cues means that a detection algorithm which utilises 3-D data is more ideal.

Time-of-Flight (TOF) camera's provide, currently, a low resolution and high frequency 3-D representation of the scene and their self-illuminating nature makes them robust to varying illumination in the environment. These desirable attributes have resulted in some techniques for people detection and tracking with the sensor. In [Tanner *et al.*, 2008] and [Bevilacqua *et al.*, 2006] a TOF camera was used for the detection and tracking of people in a static surveillance like arrangement. Similarly, [Hansen *et al.*, 2008] presented a person tracking system using a background model with a static TOF camera. To assist with tracking, the 3-D data is projected onto a 2-D which the author refers to as a *flat-map*. The person detection in these instances is not very selective. In our application, a descriptive person detection approach is essential as the platform, and consequently the sensor, is dynamic and new objects in the observed scene may not be people.

The use of 3-D data for detecting people in [Tanner *et al.*, 2008; Bevilacqua *et al.*, 2006; Hansen *et al.*, 2008] show that it is suitable for this application. The projection and clustering of the 3-D data to assist with detection in [Hansen *et al.*, 2008] is both informative and efficient. The head and shoulders were shown to provide a reliable feature to detect people with in [Li *et al.*, 2009].

In summary, the ability for robots to be aware of the location of people is a necessary piece of information for enabling HRI. The location of a person and 3-D information about the person is useful for extraction additional useful cues from a person, such as pose and gestures. Furthermore, the complex environments in which HRI occurs requires the method to be robust. Therefore, there is clearly a need for a robust person detection approach that utilises 3-D information.

This paper presents a robust person detection approach that uses 3-D range data. The breakdown of this paper is follows. First a thorough overview of the approach is provided which describes the processes involved in the locating of a person: scene segmentation, shape extraction, shape description and shape classification. Detailed results of the approach are then analysed. A database of 25 males between the ages of 20-50 was collated for testing the the accuracy of the shape matching. Finally a discussion about this method and possible future work as well as concluding remarks are provided.

2 Method

The Swissranger 4000 TOF camera has been used for this work, however the approach is suitable for sensors that can provide a 3-D representation of scene. This novel approach exploits 2-D representations of the 3-D information to facilitate efficient and robust segmentation of

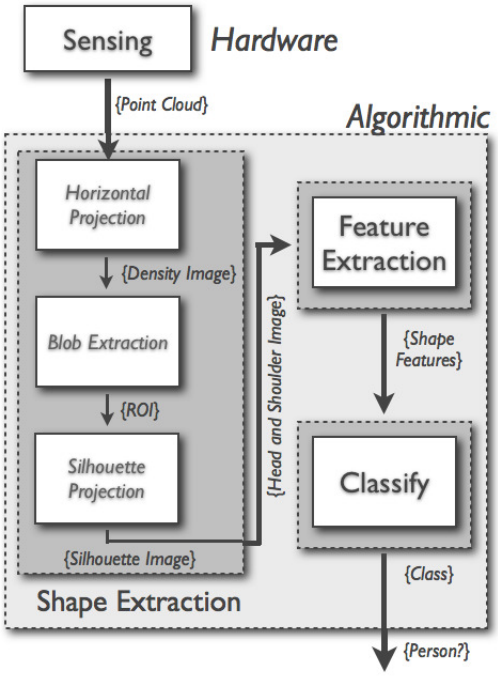


Figure 2: Process and information flow for the person detection.

the scene, whilst preserving the information necessary to locate the person. It also exploits the consistent nature of the head and shoulder region as a feature for shape matching. An important aspect of this approach, in terms of robustness and efficiency, is that it is a selective sequential method that eliminates irrelevant information at several stages. In order to be able to utilise shape matching techniques for person detection, it is important that consistent shapes are extracted from the scene for classification. Figure 2 portrays the processes that are involved in this approach.

As can be seen in Figure 2, the first step in achieving this is to segment the scene to identify objects that are potentially people. To do this, the 3-D point cloud data is projected down onto the horizontal plane, as illustrated in Figure 3, so that blob detection on the 2-D plane can be used to extract the clusters. This representation is subsequently weighted to account for the sensor characteristics, thus allowing for reliable segmentation. The 3-D information that contributes to each cluster is then projected onto another 2-D representation, the silhouette image, which is also portrayed in Figure 3. The silhouette is further segmented to extract the region that may contain the head and shoulders of a person. Shape features, FDs, are then extracted and classification with a SVM performed to determine if the shape was extracted from a person. The following section will describe each of these steps in detail.

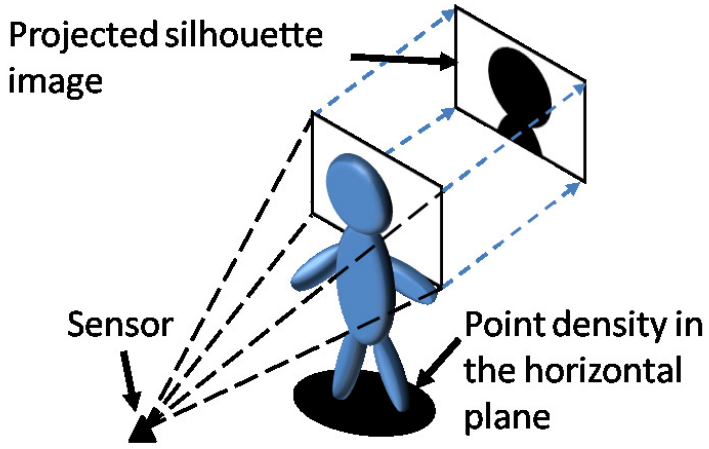


Figure 3: Illustration of the density and silhouette projections.

2.1 Scene Segmentation

This section will detail a novel approach to scene segmentation which considers the sensor properties. The first step in our approach is to segment regions in the scene that portray physical characteristics of people. The large vertical to horizontal dimensional ratio of the human form means that when a person is within the FOV of the Swissranger, the points that are returned from the person lay within a dense region when projected onto the horizontal plane, as illustrated in Figure 3. The horizontal plane is then discretised into square grids of equal size and the points above each grid are counted. This 2-D histogram, in practice, is represented by an image and will be referred to as the density image, with each pixel representing a grid. The point cloud is then projected onto this plane and for each point that lies on a grid, the intensity of the corresponding pixel in the density image is incremented by C_{inc}. Figure 4 shows the density image portrayal of the scene, where the person is clearly represented by the cluster of high intensity pixels. However, due to the focal nature of the sensor, the sensing density, D_z, decreases as a function of distance, z. This is depicted in Figure 5 and more formally specified in Equation 1. This equation illustrates how the point density is a affected by the FOV constants, θ and σ, and more significantly the depth, z.

Consequently, the same object at different distances will have different intensities in the histogram, which is undesirable for reliable segmentation. Therefore, a weighting function, Equation 2, was developed to account for diminishing sensing density resolution. The weighting function removes the negative effect that the range and FOV have on the sensing density for the purposes of consistent extraction in the density image.

The original and weighted histograms can be seen in

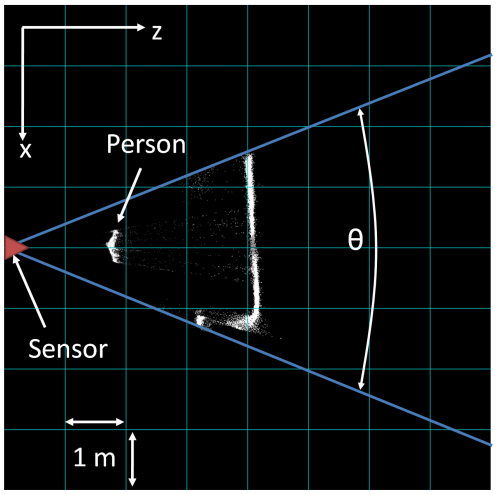


Figure 4: Annotated density image showing the axes, sensor position and FOV.

$$
\begin{aligned}
D_z &= \frac{N_{points}}{A_z} \\
&= \frac{N_{points}}{l_z \times h_z} \\
&= \frac{N_{points}}{[2 \times z \times \tan(\theta)] \times [2 \times z \times \tan(\sigma)]} \\
&= \frac{N_{points}}{4 \times z^2 \times tan(\theta) \times tan(\sigma)} \qquad (1)
\end{aligned}
$$

where:
θ is the horizontal FOV
σ is the vertical FOV
N_{points} is the sensor resolution

Figures 6(a) and 6(b) respectively. The affect that the diminishing sensing density has is clearly illustrated, as the person moves away from the sensor the intensity decreases. In the weighted histogram, the intensities of the person remain relatively constant, thus allowing for easy extraction invariant of range. The process of creating the weighted density image is shown in Algorithm 1.

The objects of interest in the scene are then extracted by thresholding the histogram and locating blobs whose sizes are representative of the physical characteristics of a person. A region-of-interest (ROI) is then defined around the centre of each object to encapsulate the cluster and consequently the information about the person. The scene segmentation process selectively reduces the large amount of raw data down to a manageable number of areas of interest, as shown in Figure 6. Information in these ROI then needs to be represented in an efficient and descriptive format for classification.

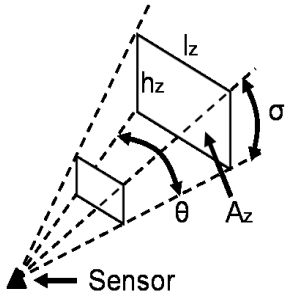
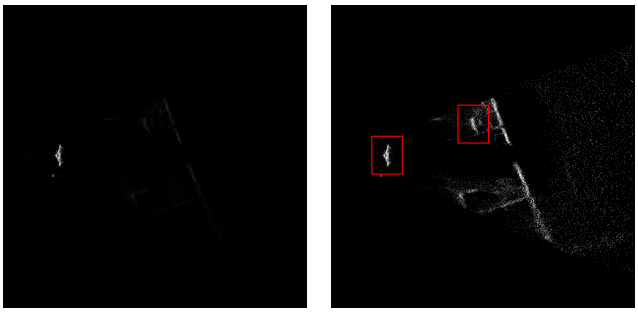


Figure 5: Illustration of the increasing sensing as a function of distance, thus illustrating how the sensing density diminishes.

$$w_z = \frac{1}{4z^2 tan(\theta) tan(\sigma)} \qquad (2)$$

where, for the Swissranger 4000 model:
$\theta = 43.6°$
$\sigma = 34.6°$



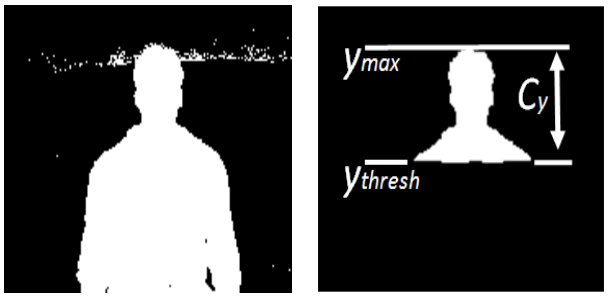
(a) Raw histogram (b) Weighted histogram with ROI

Figure 6: This figure demonstrated the affect of the density weighting function. Figure 6(b) illustrates the ROI around the clusters; one a person and the other an object in the scene.

Algorithm 1 Weighted density image

for $i = 1$ **to** N_{points} **do**
$i_x = \text{round}(P_x(i) \times h_{res})$
$i_z = \text{round}(P_z(i) \times h_{res})$
$h(i_x, i_z) = h(i_x, i_z) + (C_{inc} \times w_z)$
end for

where:
$P_x(i), P_z(i)$ are the x and z values for point i [metres]
i_x and i_z are the indices of the density image
h_{res} is the density image resolution [pixels/metre]
C_{inc} is the pixel intensity increment value



(a) Projected silhouette image (b) Extracted head and shoulders shape

Figure 7: Illustration of shape extraction process.

2.2 Shape Extraction

The head and shoulders are a descriptive yet relatively static aspect of a person. When people walk they not only move their legs, but tend to move their arms as well. This motion around the lower body and torso means that the area is highly variable in its appearance and also frequently occludes itself. The head also provides a reliable reference point as it is, typically, physically isolated from the environment. This is an advantage over other parts of the body that frequent interact with the environment, such as feet with the floor and the torso and lower body with a chair, thus making segmentation difficult. Also, many of the obstacles in the environment where humans exist, such as tables and chairs, are less likely to obstruct the sensing of the head and shoulder region. These attributes make the head and shoulder region a suitable feature for person detection using shape features.

For shape matching to be used to detect people, it is important that the head and shoulder region is robustly and consistently extracted regardless of the height or orientation of the person. Our first step in extracting this feature is projecting the 3-D information within the ROI onto a 2-D silhouette image representation. The range information is then utilised to ensure that the shape size is constant. This is achieved by first searching the corresponding range data for each positive pixel that in the silhouette image to find the highest value, y_{max}. A second value, y_{thresh}, is then calculated, as shown in Equation 3, which specifies the height threshold for the current shape. The shape is then processed to remove pixels whose corresponding height is less than this value. Figure 7(b) illustrates an example output of this process. As the threshold height, y_{thresh}, is relative to the height of the shape, y_{max}, the segmentation is invariant to the height of the person. The shape height constant, C_y, defines the size vertical size of the shape in metres, rather than pixels, thus ensuring the shape is a constant height. This is useful for generalising the solution as it

$$y_{thresh} = y_{max} - C_y \tag{3}$$

where:
y_{thresh} is the threshold height for the shape [metres]
y_{max} is the height for the shape [metres]
C_y is shape height constant [metres]

is applicable for people of varying heights and allows for scenarios when people are sitting down or similar. The constant value used for C_y is 0.4 meters as it was empirically determined to capture the head and shoulder region.

This process results in a binary shape image that is of a consistent height. Features that describe the shape now need to be extracted to allow for classification.

2.3 Shape Description

The segmentation process that is applied to the range information results in shape representations of the object of interest. This leads the task of people detection into being one of shape matching. FDs have been shown to provide good results in shape based classification tasks [Zhang and Lu, 2002]. The shape feature extraction process is illustrated in Algorithm 2, which takes a set of shapes from the aforementioned segmentation and shape extraction process and finds their Fourier descriptor features for classification.

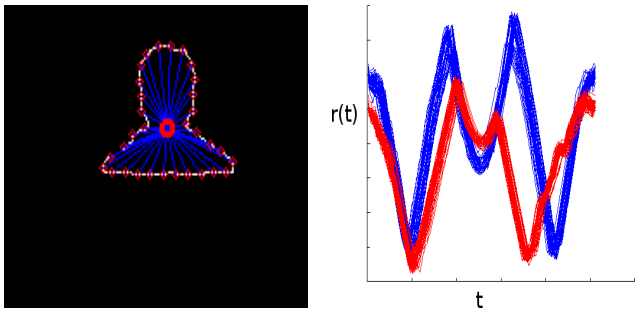
Algorithm 2 Summary of feature extraction

for $i = 1$ **to** N_{shapes} **do**
  Extract contour: $c(t)$
  Find contour centroid: $c(t) \rightarrow (x_c, y_c)$
  Sample contour: $c(t) \rightarrow s(t)$
  Calculate centroid distance vector: $s(t) \rightarrow r(t)$
  Find FDs: $r(t) \rightarrow F_k$
end for

The first step in finding the FDs is to find the contour, $c(t)$, of the shape. The contour is then sampled resulting in a reduced set of points that described the boundary, $s(t)$. This is done by sampling the contour at equal arc lengths, Figure 8(a) illustrates the sampled contour. The arc length, L, is found by dividing the perimeter, P, by the number of samples, N:

$$L = \frac{P}{N} \tag{4}$$

A mapping from the 2-D sampled closed curve, $s(t) = (x(t), y(t))$ to a 1-D representation, $r(t)$, then needs to be defined, known as the *shape signature*. There are several shape signatures that have been used to extract



(a) Sampled contour (b) Centroid Distance vs Samples

Figure 8: Illustration of the centroid shape signature. Figure 8(a) portrays the sampling of the centroid distance shape signature, $r(t)$. Figure 8(b) is a plot of the shape signature vs the sample number for 50 instances of two types of shape; blue represents the shape signature for a frontal shape and red for profile.

FDs; cumulative angle, curvature function, complex coordinates and centroid distance. The centroid distance shape signature has been shown to outperform the others in [Zhang and Lu, 2002], thus was used as the shape feature. This shape signature requires the centroid of the shape to be found:

$$x_c = \frac{1}{N}\sum_{t=0}^{N-1} x(t) \qquad y_c = \frac{1}{N}\sum_{t=0}^{N-1} y(t) \tag{5}$$

The centroid distance shape signature expresses the 2-D contour as a vector of euclidean distances between each sample point in $s(t)$, (x_t, y_t), and the centroid of the contour, (x_c, y_c), as shown in Equation 6. Figure 8(a) illustrates the extraction of the centroid distance shape signature from the contour, where $N = 32$. Figure 8(b) shows the wave form that is created by the shape signature as it moves around the shape for 50 separate instances taken from a three different people; frontal is shown in blue and profile in red.

$$r(t) = \sqrt{([x_t - x_c]^2 + [y_t - y_c]^2)},\ t = 0, 1, \ldots, N-1 \tag{6}$$

The FDs are obtained by taking the Discrete Fourier Transform (DFT) of the shape signature vector, as shown in Equation 7.

$$FD_k = \frac{1}{N}\sum_{t=0}^{N-1} r(t)e^{-j2\pi nk/N},\ k = 0, 1, \ldots, N-1 \tag{7}$$

The Fast Fourier Transform (FFT) is a computationally efficient implementation of the DFT, and is used for the calculation of the FDs. The number of sample points for the shape signature is chosen to a power-of-two for optimal efficiency of the FFT.

The centroid distance shape signature is real valued, therefore there are only $N/2$ different frequencies and only half of the FDs are required for the feature vector. For scale invariance, the magnitude of the FDs are divided by the magnitude of the DC component:

$$f_{k-1} = \frac{|FD_k|}{|FD_0|},\ k = 1, \ldots, N/2 \tag{8}$$

Now that a feature vector which describes the shape has been found, a method for determining whether or not these features represent a person is required..

2.4 Shape Classification

The final stage is to classify the feature vectors. SVMs were employed as the classifier as they are fast and have been shown to provide good classification accuracy in numerous applications. The feature vector is scaled and the scaled data is used to train the SVM [Chang and Lin, 2001] offline using five fold cross validation. Different models were trained so that the accuracy of the class separation for different shape features could be analysed.

3 Results

The following section details the results of the empirical evaluation of the presented approach. First, the set-up of the experiment from which the results are drawn from are discussed. The result of applying the weighting function to the density image and the shape extraction process, in various scenarios, are then outlined. Finally, a thorough analysis of the shape classification process is provided.

3.1 Experiment Set-up

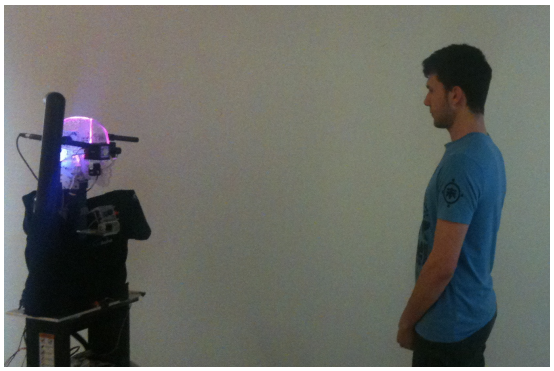


Figure 9: Illustration of experiment set-up.

A TOF camera database consisting of two different scenarios was collected with the robotic platform, the

Table 1: Distribution of training and testing data

Classes	Train Samples	Test Samples	Total
Negatives	457	825	1282
Frontal	301	680	981
Profile	371	1836	2207

first of the two scenarios is shown in Figure 9, for testing, training and evaluation of the approach. The database consists of 25 male subjects between the ages of 20-50. The scenarios involved were:

- **Frontal:** standing facing towards the robot, as shown in Figure 9
- **Profile:** standing side on in front of the robot

In addition to these datasets, a collection of shapes that do not represent a person were collected using the Swissranger so that a negative class could be constructed. These datasets were utilised to train and test different SVM models. An overview of how the data was distributed for this is shown in Table 1. The frontal and profile training data was taken from the first 6 subjects and tested on the remaining 19, the training data was not used for testing.

3.2 Scene Segmentation

The result of applying the weighting function to the density image will now be outlined. In the raw density images, Figures 10(a) and 10(b), the intensity of the person is shown to decrease as they move away from the sensor. When the weighting function is utilised, Figures 10(c) and 10(d), it can be seen that the intensity of the person remains constant as they move away, thus allowing for reliable extraction using blob detection.

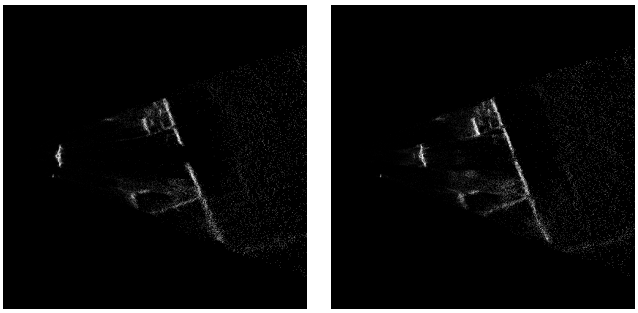
3.3 Shape Extraction

The ability to extract the head and shoulder shape for shape matching will now be shown. Figure 11 shows the successful extraction of this region for both frontal, Figure 11(a), and profile, Figure 11(b), scenarios. Furthermore, it was able to successfully extract head and shoulder of each of the subjects in the experiment.

3.4 Analysis of Classification Models

To analyse the separable nature of the features, three different SVM classification models were trained: negative vs frontal, negative vs frontal and profile, and negative vs frontal vs profile. The following will present the classification results for these models on the testing data.

(a) Raw Histogram at $slice_{t1}$ (b) Raw Histogram at $slice_{t2}$



(c) Weighted Histogram at $slice_{t1}$ (d) Weighted Histogram at $slice_{t2}$

Figure 10: Histogram representations of the scene at two time slices. 10(a) and 10(c) portray the first time instance for the raw and weighted histograms. The cluster of high intensity points to the left of the histograms represent the person. The person then moves away from the sensor as is shown in 10(b) and 10(d).

(a) Frontal (b) Profile

Figure 11: Example shape extraction results from the experiment dataset.

Negative vs Frontal

The two class negative vs frontal model was tested to see how robustly these classes could be separated, thus allowing for reliable detection of people front on. Table 2 shows the confusion matrix for the testing of the model. It is evident that the classifier is able to separate these classes well, correctly labeling the negative and frontal test data with 99.27% and 99.7% accuracy respectively.

Table 2: Confusion matrix for Negative vs Frontal model

Test Data	Negative	Frontal
Negative	819	6
Frontal	2	678

Negative vs Frontal and Profile

In the negative vs frontal and profile model, the frontal and profile features are combined to make a single class, thus it is also a two class SVM model. Table 3 shows the results of the testing data in a confusion matrix. Although the accuracy of the frontal and profile class is high, 96%, the accuracy of the negative class is reduced to 80% and a significant number of negatives are labeled as a person. This suggests that the combined frontal and profile class may be too general.

Table 3: Confusion matrix for Negative vs Frontal and Profile model

Test Data	Negative	Frontal and Profile
Negative	660	165
Frontal and Profile	101	2415

Negative vs Frontal vs Profile

The combined frontal and profile class is now split to form a negative vs frontal and profile model. The confusion matrix for the testing of this model are shown in Table 4. The accuracy of the negative class is improved significantly to 92.6% over the negative vs frontal and profile model. The accuracy of the frontal class, 99.4%, is very close to its accuracy in the negative vs frontal only model. The 94.2% accuracy of the profile class is also very good. It is evident from the confusion matrix that the greatest confusion is between the negative and profile classes which indicates that the profile shape feature is less discriminate then the frontal.

Table 4: Confusion matrix for Negative vs Frontal vs Profile model

Test Data	Negative	Frontal	Profile
Negative	764	4	105
Frontal	4	676	0
Profile	105	1	1730

The classification results show that the novel shape features are a good representation of a person as they were able to discriminatively detect a person. Furthermore, the ability to distinguish between the frontal and profile orientation of a person has been displayed, which is a useful information in HRI applications.

4 Discussion

To reiterate, although a TOF camera was used for this implementation of the person detection technique, the shape segmentation, feature extraction and classification process are a generic solution provided a 3-D point cloud can be obtained. However, the invariance to illumination that the TOF camera provides is ideal for this application as it allows for extreme changes in illumination, i.e. the lights turned on or off. It must be noted that TOF cameras suffer from a phase unwrapping problem. This means that objects which are outside the maximum sensing range of the sensors, the ambiguity range, will provide erroneous readings as the phase shift can not be accurately unwrapped. There have been methods proposed in the literature to resolve this issue and need to be investigated further. However, as was stated previously, this approach is not sensor specific and requires only 3-D data and fixing this sensor is not the focus of this paper.

5 Conclusion and Future Work

A shape based person detection approach that uses 3-D data has been presented. The use of 2-D projections of the 3-D data for processing has resulted in a computationally efficient method to segment the scene and extract the shape information which is used to detect a person. The results have shown that objects can be segmented and the head and shoulder region extracted in various poses and with different people. The classification results show that this approach can, with high accuracy, discriminate between a person and an object from the environment, which is valuable in HRI as it can work in cluttered environments; such as the home. Furthermore, it was shown that the classification can differentiate between frontal and profile poses of the person. This is a useful attribute for HRI as it not only allows for the robot to understand where the person is in the scene, but also a coarse pose for the person.

Currently, the database consists of adult males in frontal and profile poses and therefore future work is required to expand this database so that the classification for different poses and groups, such as female and children, can be verified. It would also allow for the possibility to explore the extension of this method to differentiate different groups; e.g. children vs adults. Furthermore, the potential to extend this approach so that pose and gestures can be extracted will be investigated.

6 Acknowledgment

This work is supported by the ARC Centre of Excellence for Autonomous Systems, RobotAssist and the University of Technology, Sydney.

References

[Bertozzi *et al.*, 2007] M. Bertozzi, A. Broggi, M. Del Rose, M. Felisa, A. Rakotomamonjy, and F. Suard. A pedestrian detector using histograms of oriented gradients and a support vector machine classifier, 2007.

[Bevilacqua *et al.*, 2006] Alessandro Bevilacqua, Luigi Di Stefano, and Pietro Azzari. People tracking using a time-of-flight depth sensor. In *AVSS '06: Proceedings of the IEEE International Conference on Video and Signal Based Surveillance*, page 89, Washington, DC, USA, 2006. IEEE Computer Society.

[Buciu *et al.*, 2001] I. Buciu, C. Kotropoulos, and I. Pitas. Combining support vector machines for accurate face detection. In *In Proc. of ICIP01*, pages 1054–1057, 2001.

[Cezar Silveira Jacques *et al.*, 2006] J. Cezar Silveira Jacques, C. Rosito Jung, and S.R. Musse. A background subtraction model adapted to illumination changes. pages 1817 –1820, oct. 2006.

[Chang and Lin, 2001] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005.

[Hansen *et al.*, 2008] D.W. Hansen, M.S. Hansen, M. Kirschmeyer, R. Larsen, and D. Silvestre. Cluster tracking with time-of-flight cameras. pages 1 –6, jun. 2008.

[Li *et al.*, 2009] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Rapid and robust human detection and tracking based on omega-shape features. In *ICIP'09: Proceedings of the 16th IEEE international conference on Image processing*, pages 2517–2520, Piscataway, NJ, USA, 2009. IEEE Press.

[Rowley *et al.*, 1998] Henry A. Rowley, Student Member, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions On Pattern Analysis and Machine intelligence*, 20:23–38, 1998.

[Tanner *et al.*, 2008] Rudolf Tanner, Martin Studer, Adriano Zanoli, and Andreas Hartmann. People detection and tracking with tof sensor. In *AVSS '08: Proceedings of the 2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*, pages 356–361. IEEE Computer Society, Washington, DC, USA, 2008.

[Viola and Jones, 2004] Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[Waring and Liu, 2005] Christopher A. Waring and Xiuwen Liu. Face detection using spectral histograms and svms. *IEEE Trans Systems, Man, and Cybernetics-Part B: C Cybernetics. 2005*, 35:467–476, 2005.

[yu Lin, 2005] Yen yu Lin. Robust face detection with multi-class boosting. In *In CVPR 2005*, pages 680–687, 2005.

[Zeng *et al.*, 2008] Hui-Chi Zeng, Szu-Hao Huang, and Shang-Hong Lai. Real-time video surveillance based on combining foreground extraction and human detection. In *MMM'08: Proceedings of the 14th international conference on Advances in multimedia modeling*, pages 70–79, Berlin, Heidelberg, 2008. Springer-Verlag.

[Zhang and Lu, 2002] Dengsheng Zhang and Guojun Lu. A comparative study of fourier descriptors for shape representation and retrieval. In *Proc. of 5th Asian Conference on Computer Vision (ACCV*, pages 646–651. Springer, 2002.