A robust people detection, tracking, and counting system

Nathan Kirchner[1], **Alen Alempijevic**[1], **Alexander Virgona**[1]
Xiaohe Dai[2],**Paul G. Plöger**[3], **Ravi Kumar Venkat**[3]
[1]Centre for Autonomous Systems, University of Technology, Sydney, NSW Australia
{nathan.kirchner, alen.alempijevic, alexander.virgona}@uts.edu.au
[2]Institue of Cyber-Systems and Control, Zhejiang University,Hangzhou, China
dai@iipc.zju.edu.cn
[3]Bonn-Rhein-Sieg University of Applied Science, Sankt Augustin, Germany
paul.ploeger@h-brs.de, ravi.venkat@smail.inf.h-brs.de

Abstract

The ability to track moving people is a key aspect of autonomous robot systems in real-world environments. Whilst for many tasks knowing the approximate positions of people may be sufficient, the ability to identify unique people is needed to accurately count people in the real world. To accomplish the people counting task, a robust system for people detection, tracking and identification is needed.

This paper presents our approach for robust real world people detection, tracking and counting using a PrimeSense RGBD camera. Our past research, upon which we built, is highlighted and novel methods to solve the problems of sensor self-localisation, false negatives due to persons physically interacting with the environment, and track misassociation due to crowdedness are presented. An empirical evaluation of our approach in a major Sydney public train station (*N=420*) was conducted, and results demonstrating our methods in the complexities of this challenging environment are presented.

1 Introduction

Multi-person tracking plays an important role in dynamic environment understanding and is crucial to the operation of robots in the real world. Whilst for some traditional problems, such as path planning and collision avoidance, it is sufficient to know the positions of people, for problems where we need to know the density of people it is important to maintain the complete trajectories of each person and identify them through a scene.

To date, various approaches have been explored for person detection, tracking, and/or counting by a number of researchers. Several representative examples are [Zhao *et al.*, 2009], [Garcia *et al.*, 2013], and [Chen *et al.*, 2012]. Whilst these approaches have shown promise, their core assumptions limit their suitability for our target problem; where people will be moving in all directions (including away from the sensor), may be wearing hats and/or glasses, may appear in numbers in any one observation, and foreseeably may not directly face the sensor at anytime as they pass. For instance, [Zhao *et al.*, 2009] detects persons in images of the scene via face detection, and continues to track and count people (reported accuracy of 93%). However, a prerequisite of this approach is that faces will be observable, which is foreseeably not the case in our problem. The head-based person detection presented in [Garcia *et al.*, 2013] moves past this particular limitation. However, the authors acknowledge that the reported 98% overall accuracy will be significantly impacted with increasing numbers of people in single observations and with occlusions, both of which are common to our problem.

In [Chen *et al.*, 2012] person detection is achieved through image processing which targets more general features. Tracking and counting is then achieved through analysing bounding box behaviour between successive frames. However, limitations were again reported that are expected in our problem; specifically, path behaviour such as turning and/or moving quickly.

The inherent properties of a 2D sensed data based approach have been noted by many researchers, and a number of approaches have been built on 3D sensed data. In the stereo camera based approach of [Qiuyu *et al.*, 2010] the camera is mounted to produce top-down observations and objects are recognised using disparity maps. The author report impressive results that are robust to illumination variations and crowding. However, the approach requires particular placement of the sensor which precludes use in areas such as typical passageway which generally have lower ceilings. Furthermore, this approach can not differentiate between a person and a moving object such as trolley-luggage; a foreseeable limitation in our problem.

The Time of Flight sensor approach of [Hsieh *et al.*, 2012] while showing promise in several respects (such as crowded scenes and partial occlusions) shows limitations in the differentiation of a person from a moving object. The approached presented in [Zhu and Wong, 2013] moves past this limitation through using depth images to explicitly detect people. However, stated limitations include cases in which multiple people are close and/or are in physical contact; which are foreseeable in our application space.

From this, it seems that approaches for person detecting, tracking, and counting for a set of environments and/or assumptions exist. However, it seems that they have limited suitability for our real world scenario. An approach that is capable of the non-trivial task of robust person detection, tracking and counting in real world scenarios is required. Specifically, the approach must be robust against sensor positioning, the plane of observation of sensed persons, their variation in speed and direction, pose and physical interactions with the environment and each other, feature occlusions or truncation by the sensor's field of view, moving objects, periods of no observations, and must be capable of distinguishing unique persons, and handling multiple people in the field of view.

This paper presents an overview of our approach; our past research upon which we built is highlighted and our novel contributions related to this publication are detailed. Specifically, this paper details our efforts building on our past robust person detection research [Hordern and Kirchner, 2010]. There are four main novel contributions of this work. First, a framework that encapsulates people detection, tracking and counting was devised. Secondly, a method to enable sensor self localisation was devised. Thirdly, a method for exploiting environment information and Australian building standards to provide robustness against persons physically interacting with the environment was devised. Finally, a method for exploiting the people's (Head-to-shoulder signature - HSS), which is introduced in our previous work [Kirchner *et al.*, 2012], that enabled unique identification of people was devised and shown to be effective in a track merge/split event.

The breakdown of this paper is as follows: Section 2 provides background on our approach. Section 3 details the methods through which these contributions were made. An empirical evaluation in a public train station (*N=420*) is presented in Section 4. Conclusions are drawn, limitations identified, and future work is proposed in Section 5.

2 Background

This section details our previous work which forms the foundation of the contributions of this paper. Firstly, our previous efforts towards person detection are described. Following which, our work towards tracking is detailed.

Our robust person detection is a three step process: namely 1) point cloud segmentation, 2) feature vector construction, 3) classification. The segmentation stage, originally presented in [Hordern and Kirchner, 2010], reduces the search space by identifying and segmenting sub-areas of the input 3D point cloud. This stage creates a *density image* through projecting the 3D points onto the on the horizontal plane and conducting a bivariate histogram. This process inherently results in relatively larger numbers of points being projected onto relatively smaller numbers of histogram bins in the case of more vertical surfaces in the point cloud; whereas more horizontal surfaces result in relatively smaller numbers of points being projected onto relatively larger numbers of histogram bins.

The number of points present in *bin a* at distance x for a horizontal surface is estimated and used as a lower-threshold to identify bins with surfaces of interest. The three levels (*Red*, *Orange*, *Green*) represent the belief of whether a blob is a person or not. With some prior knowledge, we know that a person blob is usually more than 0.1m in width and less than 1m in length.

Groups of adjoining bins are checked for a minimal size >0.1m, if true the person detect level is *Red*. The blobs are then checked for a maximum of <1m, if true the person detect level is upgraded to *Orange*. These dimension checks remove blobs from larger surfaces such as walls and/or geometric blobs that are not human-shape. This approach exploits the reasonable assumptions that people tend to appear as vertical surfaces in the point cloud. This process is susceptible to false positives, however it has been demonstrated acceptable by our previous work as discrimination is primarily the function of the feature vector / classification stage. Whilst an approach for stage 2) and 3) was presented in [Hordern and Kirchner, 2010], this approach was improved in [Kirchner *et al.*, 2012] and that approach is adopted here.

This method further exploits the generalised size, shape, and subsequent ratios of people's head-to-shoulder region for person detection through constructing a scale and viewing angle robust feature vector (HSS). Lateral measurements of the head-to-shoulder region are explicit. HSS are constructed by taking horizontal slices of the 3D point cloud segmented by stage 1, calculating the span of each slice, and collating those into a feature vector. These spans, derived from 3D data, are relatively consistent for the 360^o range of observation angles - other than perfect-portrait and perfect-profile where shadowing is significant. The vertical characteristics are not explicitly measured, but are encapsulated in the HSS via the feature vector location in which particular lateral measurements appear. The HSS feature vectors are the basis upon which a two class trained Support Vector Machine (SVM) is used to detect people.

The SVM was trained with supervised data of 10 different people for the person class, and a number of post stage 1 segmentation data known to not be a person. The person detect level for blobs classified as people by the SVM are upgraded to *Green*.

The results presented in [Hordern and Kirchner, 2010] and [Kirchner *et al.*, 2012] demonstrate the robustness of our person detection approach against persons moving speeds, directions and pose, orientation relative to the sensor (including facing directly away), typical occlusions from people crowding, and against false positives from objects moving and otherwise. However, stated limitations include the requirement for the sensor positioning and plane of observation of sensed persons to be known, and that persons physically interacting with the environment are not detected.

Having detected people, a particle filter is used in order to perform global tracking as per our previously published work [Alempijevic *et al.*, 2013]. To estimate each positional trace, a set of samples $X_t = \langle x_t^i \mid i = 1...N \rangle$ and its associated weights ω_t^i representing the belief at time t of the persons location are generated. The computation of the posterior for each t-th particle set X_t is then calculated recursively from X_{t-1} in three steps as detailed in Alg.1. And a Gaussian random motion model is used for prediction of motion $u_t = \mathcal{N}(\mu, \sigma)$, where μ is 1.4m/s and σ 0.5 [Browning *et al.*, 2006].

```
Algorithm 1: Particle filter algorithm key steps
    Prediction: Draw $x_t^i \sim p(x_t^i \mid x_{t-1}^i, u_{t-1})$.
    Update: Compute weights $\omega_t^i = \eta p(y_t \mid x_t^i)$, with $\eta$
    a normalisation factor to so that the weights sum to
    1. Here, $y_t$ is a person detection reading at time $t$.
    Resample: Draw new $x_t^i$ using weights $w_t^i$.
```

3 Method

This section presents our approach to people detection, tracking and counting. Our previous work is placed within the overall explanation, and the novel methods through which the contributions of this paper were realised are detailed. Fig. 1 presents the framework of our approach. *Self Localisation*: this novel stage contributes robustness against sensor positioning by enabling our sensing platform to self localise upon installation. *People Detection*: this stage is built on our previous work described in Section 2. However, a novel element has been contributed to address the ubiquitous issue of false negatives due to persons physically interacting with the environment. *People Tracker*: again, a novel element has been added to our previous work to contribute robust individual-specific features based track association to resolve ambiguities. *People Counter*: this stage exploits the inherited robustness from previous stages to count people.

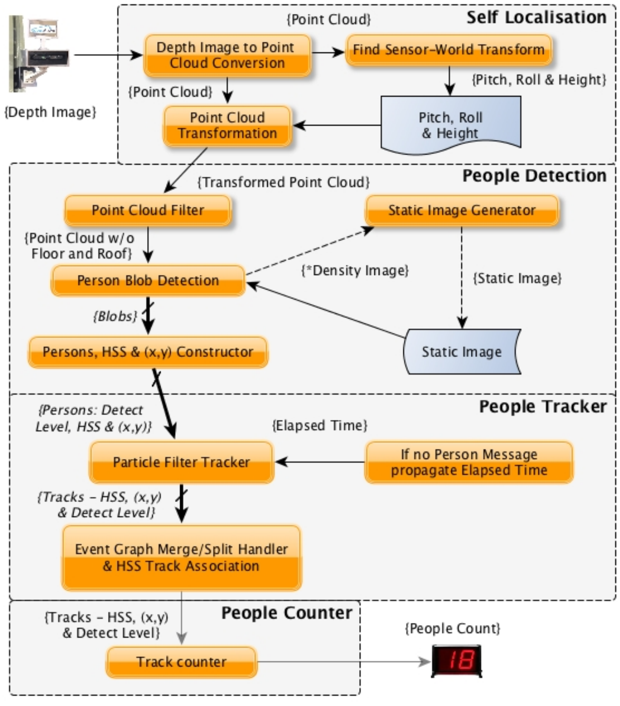

Self Localisation
{Point Cloud}
{Depth Image}
Depth Image to Point Cloud Conversion
Find Sensor-World Transform
{Pitch, Roll & Height}
{Point Cloud}
Point Cloud Transformation
Pitch, Roll & Height
People Detection
{Transformed Point Cloud}
Point Cloud Filter
Static Image Generator
{Point Cloud w/o Floor and Roof}
Person Blob Detection
{*Density Image}
{Static Image}
{Blobs}
Persons, HSS & (x,y) Constructor
Static Image
People Tracker
{Persons: Detect Level, HSS & (x,y)}
{Elapsed Time}
Particle Filter Tracker
If no Person Message propagate Elapsed Time
{Tracks - HSS, (x,y) & Detect Level}
Event Graph Merge/Split Handler & HSS Track Association
People Counter
{Tracks - HSS, (x,y) & Detect Level}
Track counter
{People Count}
18

Figure 1: Framework of our system

3.1 Self Localisation

The sensor frame relative to the world frame must be known in order to transform the positions of sensed persons in world frame. Due to real-world constraints, it is not always possible to manually determine the position of the sensor. As such, a necessity for sensors self localisation arises.

The Self Localisation stage is inspired by our previous research in horizontal surface detection as a prerequisite for assistive robot object manipulation [Caraian and Kirchner, 2010]. This research proposed, tested, and evidenced that the assumption that horizontal surfaces would result in relatively large collections of points on an isolatable plane holds and is robust in the real world. This assumption is exploited for a solution to the inverse problem; where the sensor-world transform is not known, but the presence of a large horizontal surface (the floor) is ensured during platform installation.

Self Localisation begins with conversion of an acquired depth image to a 3D point cloud in the sensor coordinate frame. Our method for finding the ground and determining the pitch, roll and height of the sensors is detailed in Alg. 2. Firstly, we initialise a plane using one point from the point cloud and then obtain the neighbouring points of the original point through region growing [Rabbani *et* *al.*, 2006]. Neighbouring points are checked to determine if they lie in the same plane as the original point. If they do, the parameters of the plane are updated. This is reiterated for all points in the point cloud. Finally, we will get all the planes in the point cloud. We will take the largest plane as the ground and get the formulation of the plane and calculate the pitch, roll and height of the sensor relative to plane.

Algorithm 2: Steps of Sensor Self Localisation

```
Input: $P \longleftarrow$Current point cloud , $N \longleftarrow$normals,
       $r \longleftarrow$residuals, $\Omega$neighbour finding function,
       $\theta_{th}$residual threshold
Output: pitch, roll, height from sensor origin to
        ground plane
Floor Detection:
Region List $\{\mathbf{R}\} \leftarrow \phi$,
Available points list $\{\mathbf{A}\} = \{1 \cdots P_{count}\}$
while $\{\mathbf{A}\}$ is not empty do
    Current region $\{\mathbf{R_c}\} \leftarrow \Phi$, Current seeds
    $\{\mathbf{S_c}\} \leftarrow \phi$
    Point with minimum residual in $\{\mathbf{A}\} \rightarrow P_{min}$
    Insert $P_{min}$ to $\{\mathbf{S_c}\}$ & $\{\mathbf{R_c}\}$
    Remove $P_{min}$ from $\{\mathbf{A}\}$
    for $i = 0$ to size($\{\mathbf{S_c}\}$) do
        Find nearest neighbours of current seed
        point $\{\mathbf{B_c}\} \leftarrow \mathbf{\Omega}(S_c)\{\mathbf{i}\}$
        for $j = 0$ to size($\{\mathbf{B_c}\}$) do
            Current neightbor point $P_j \leftarrow B_c\{\mathbf{j}\}$
            if $\{\mathbf{A}\}$ contains $P_j$ and
            $cos^{-1}(|\langle \mathbf{N}\{\mathbf{S_c}\{i\}\}, \mathbf{N}\{P_j\}\rangle|) < \theta_{th}$ then
                Insert $P_j$ to $\{\mathbf{R_c}\}$ Remove $P_j$ from
                $\{\mathbf{A}\}$
                if $r\{P_j\} < r_{th}$ then
                    Insert $P_j$ to $\{\mathbf{S_c}\}$
    Add current region to global list $\{\mathbf{R_c}\}$ to $\{\mathbf{R}\}$
Sort $\{\mathbf{R_c}\}$ according to the size of the region
Floor = the largest region
Localise: Get equation of plane, calculate pitch,
roll, height from plane equation
```

As a function of this process, the *height* of the sensor relative to the floor is available. Thus, at this point the determined *pitch*, *roll*, and *height*, values are passed to the Point Cloud Transformation block. Newly arriving data can then immediately pass through using this transform and can bypass the relatively slow Find Sensor-World Transform block.

3.2 People Detection

Preliminary explorations made evident that people often physically interact with the public transport environment - for instance, passengers hold hand rails while on stairs. As such, addressing this limitation of our previous work approach was a necessity.

The Point Cloud Filtering block takes the data in the global frame passed by Self Localisation stage. To reduce the computational expense, we first reduce the search space. This reduction, typically ≈50%, is achieved through removing the floor/stairs and ceiling points. With the floor plane's height known from the sensor self localisation stage, stairs are removed using the fixed size/gradient prescribed by prior knowledge of staircase gradient from Australian Standards, and from the onset of the stairs. The ceiling plane is removed using a threshold 2.2m, because the height of people is generally less than 2.2m.

The Person Blob Detection block automatically finds a no-person frame and captures a background (static image). The background contains information of static environment factors such as handrails. Subsequently this static image is used by our system to negate the background from each bivariate histogram obtained. As such, people are separated from the facilities with which they physically interact.

The remaining blocks of this stage, including HSS feature vector construction, are as per our previous research; detailed in Section 2 and in [Hordern and Kirchner, 2010] and [Kirchner *et al.*, 2012]. The outputs of the People Detection stage are individual data structures for each detected person that carry that person's detect level, $\{x,y\}$ position and their associated HSS (as shown in Fig. 1).

3.3 People Tracker

Having extracted persons and determined individual HSS, the next necessity was to track individuals while moving in the sensor field of view amidst person-person interactions, frequent changes of direction and occlusions due to permanent infrastructure. Person tracking is achieved using Bayesian multi-target tracking, an independent particle filter is associated with each individual track, capturing the location of all targets given all observations. The particle filter offers a degree of robustness to unpredictable motions, nonlinear and non Gaussian measurements.

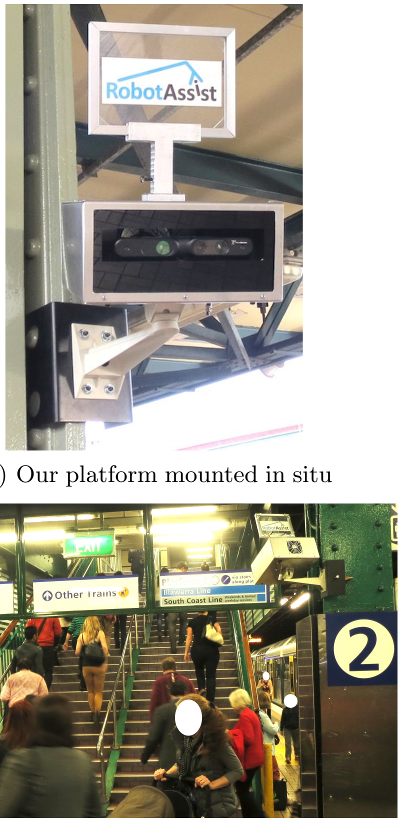
The Tracker manages track initiation, track validation and track deletion using the observations created by the Person Detection stage and the Bayesian formulation of the particle filter. A track is initialised with all person detect level *Red* observations, tracks are validated once an associated observation contains a HSS with a detect level of *Green*. Finally, tracks are deleted once the uncertainty of the associated particle filter exceeds 0.2m (≈ radius of a person's torso).

In traditional computer vision the target appearance is hard to model, thus, modelling the interaction of targets is generally required. As observations related to individuals are available (HSS), tracking the identity of targets during interactions is achieved with an Event Graph Track Handler and subsequent appearance model evaluation. The Event Graph handler with subsequent data association has been demonstrated working with 3D Lidar correlated RGB data in [Morton *et al.*, 2013]. The approach presented by the authors exploits a joint colour histogram in the HSV space for identify reasoning. This can not work in typical environments we are examining, where the clothing is consistent in large groups. Therefore, we exploit the HSS associated with individual tracks to establish the identify of each person post merge/split events.

The Event Graph handler merges and splits tracks based on the mean distance for Personal Zone interaction 0.8 m [Hall, 1966]. If a person track disappears within the distance of personal interaction from another person, this is treated as a merge event. The individual track IDs are associated to one common track and the respective HSS stored for identity reasoning. If a track (person) appears within the distance of personal interaction of an already merged track the HSS signature is used to validate the identify of each person post this split event with the stored HSS signatures. This Track Association is based on the sum of residuals between the current HSS and HSS associated with the person prior to the merge event being less than 4mm per slice (related to sensor noise). Finally, tracks that contain the corrected IDs reflecting the merged or post split created tracks are published.

3.4 People Counter

The People Tracker stage was sufficiently robust to allow reliance on track IDs as an indication of known persons. This function has been placed as a discrete stage in our framework to house potential future needs based developments.



(a) Our platform mounted in situ

(b) A typical scene during evaluation (Our platform is top-right)

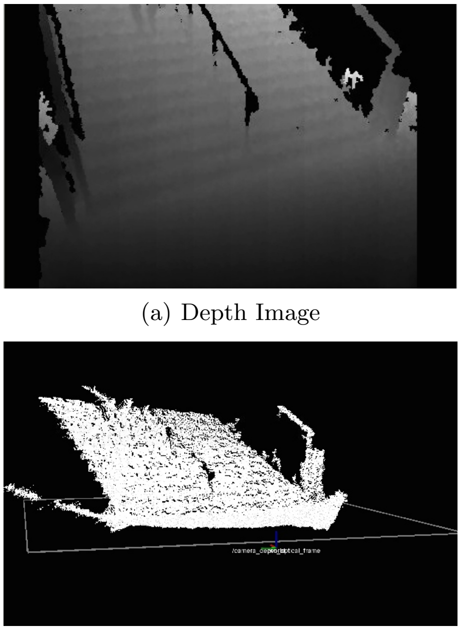
Figure 2: Our platform is primarily instrumented with a PrimeSense depth camera and has been engineered for mounting to typical public transport environment infrastructure.

4 Empirical Evaluation

Experimental evaluation was conducted at a major Sydney public train station. Fig. 2 shows our platform and an image of a scene during evaluation. Our platform consists of a PrimeSense RGBD camera and a Fit-PC3. We exploited mechanical interference with the flange rivets on the common 'vertical' structural I-beams, shown in Fig. 2. Our platform was installed in six locations and depth images were collected in two periods for each location. The evaluation was conducted from ≈ 9am-11am on a weekday. Four locations were on train platforms, the remaining two were on the concourse. In all locations our platform was mounted similarly to that shown in Fig. 2. The passenger crowdedness fluctuated with the train services from sparse (one new passerby appearing every ≈22s) to more populated (one every ≈0.4s). The data includes typical passenger egress around the station (*N=420*).

Our evaluation began with an exploration targeted at the Self Localisation stage, after the self localisation stage, the system will convert the point cloud from the local frame into the global frame as shown in Fig. 3.

In order to confirm our approach's ability to handle crowded scenes a ≈10s period of data was manually extracted in which up to 9 complete head-to-shoulder regions of in-motion people appeared simultaneously in the field of view. Fig. 4b) shows one depth image from this data. Fig. 4a) shows an RGB image from that corresponds in time with the depth image; note however, that the two images are not registered and do not have equivalent fields of view. The RGB image is included for the readers visual reference only. The instantaneous person count from our approach was compared to a manually derived instantaneous person count ground truth



(a) Depth Image

(b) After Sensor Calibration

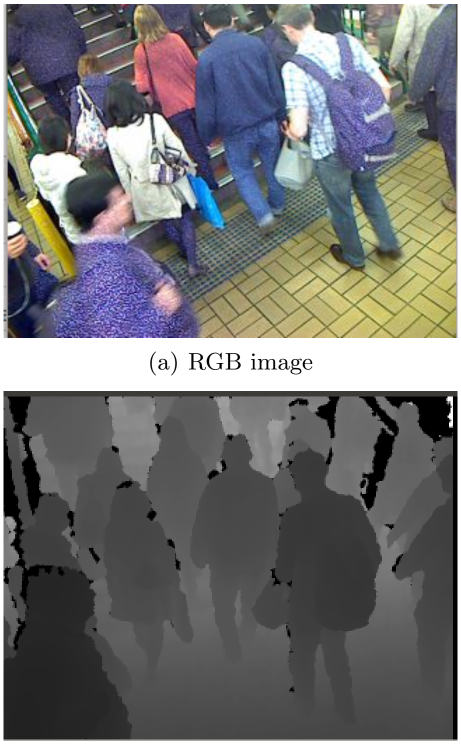
Figure 3: An example of online sensor calibration

on a frame-by-frame basis. The two counts were found to be consistent indicating our approach's ability to handle crowded scenes.

Fig. 5 shows an example of a person physically interacting with the environment, in this case the common occurrence of holding the stairs' hand rail. As can be seen in the density image of Fig. 5a) the 'L' shaped blob created during the person detect stage is red as it failed the previously mentioned constraint checks. This is due to the physical interaction with the environment, the person has joined a part of the environment and the blob reflects this. However, as can be seen from Fig. 5b) our static image side-chain effectively removed the environment, and the human-shaped blob (orange) passed this check and continued through the person detect process. Such occurrences were prevalent in our data and in all cases our static image side-chain resulted in the people becoming detectable.

Our approach handles cases where multiple people are close and/or are in physical contact through the previously described Merge/Split Handler and the subsequent HSS-based Track Association. Fig. 6 shows an example of such an event; the top row of the figure shows an RGB image sequence and the bottom row shows plots of the corresponding traces produced by our tracker. Of note from the RGB images is that these two individuals are wearing similarly coloured clothes (an occurrence observed to be common in our data); the track association of [Morton *et al.*, 2013] would likely fail.

As can be seen, two people enter the bottom-left of the



(a) RGB image

(b) Depth image

Figure 4: An example scene from the extracted data segment. Note: the two images are not registered and do not have equivalent fields of view.

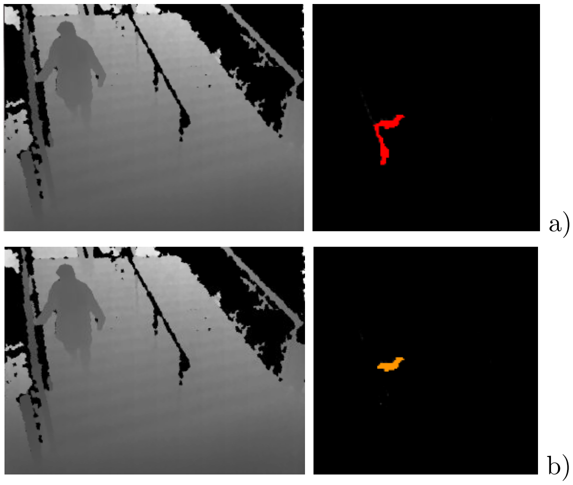


Figure 5: A person detection example - a depth image and corresponding density image a) without and b) with our static image side-chain.

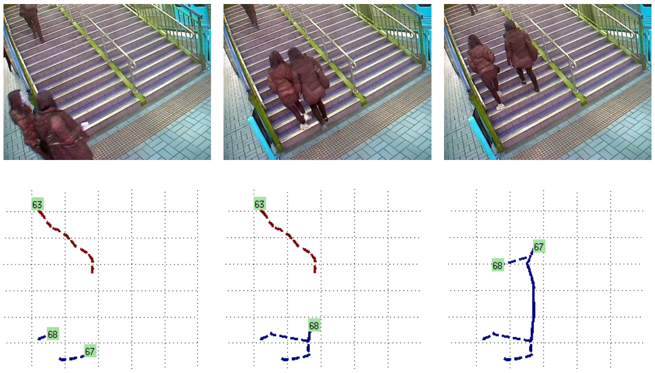


Figure 6: An image sequence of an Merge/Split event. RGB images (top row) and trace plots (bottom row) of this event are shown.

Table 1: Person count performance of our approach

	Counts			
S.No.	Manual	Software	FPos	FNeg
09-27-04	33	32	0	1
09-31-27	18	17	0	1
09-33-33	72	63	0	9
09-40-24	43	42	0	1
10-02-24	11	8	0	3
10-18-47	18	16	0	2
10-22-16	28	23	0	5
10-30-28	25	12	0	13
10-33-10	41	33	0	8
10-38-03	33	31	0	2
10-53-26	98	86	0	17
Total	420	372	0	48

image as individuals, they move very close to each other and merge, and then again separate. The bottom row of the figure shows that our Merge/Split Handler and subsequent HSS-based Track Association correctly handles the event. Specifically, two traces (shown with dotted lines and numbered 67 & 68) begin upon detection of the two individuals. A merge event is detected (second image) and a hybrid trace (shown with solid line and numbered 68) begins. In the third image a split event is detected and the two traces are successfully associated with their original track IDs (shown with dotted lines and numbered 67 & 68). As HSS-based person discrimination has previously been demonstrated reliable and robust [Kirchner *et al.*, 2012], discrimination was not explored extensively here. However, the in-practice association performance was verified over 5 randomly selected merge/split events.

Finally, in order to explore the performance of our approach the depth images from all 12 collection periods were manually reviewed with the appearance of each person noted, and the total calculated. The manual review was conducted by one researcher who independently reviewed each data period several times, and on separate occasions. These results were validated by four other researchers who each selected several periods for independent review. Inconsistencies were then re-reviewed. The criteria for 'detecting a person' was the time at which the entire head-to-shoulder region appeared in the field of view. This data was then used with our approach, implemented as described in Section 3, with the total count noted. These results are shown in Table 1.

As can be seen from the table our approach successfully detected and counted people without false positives (a critical design requirement). As our past work [Kirchner *et al.*, 2012] has demonstrated that the reliable range for person detection is 4m, data beyond this distance is excluded by the person detector. As such, the data was cropped at 4m before being used to produce the manual persons counts. A number of false negatives are however evident. Manual case review of these suggest that they are due to extreme truncation of persons in the sensor's field of view, which is especially evident when the sensor was placed on the open area of a concourse (data file 10-53-26). In which case the HSS feature vector is malformed and can not pass the SVM classification.

Nevertheless, these results clearly show that our methods described in Section 3 significantly add to our approach for real world scenarios such as this public train station. Furthermore, these results serve to highlight the additional complexity and required contributions for devising externally valid and robust methods.

5 Conclusion and Future Works

This paper presented our approach for robust real world people detection, tracking and counting. Our past research, upon which we built, was highlighted and our novel contributions specific to this paper were detailed. Specifically, our methods for adding robustness against sensor positioning, false negatives due to persons physically interacting with the environment, and track misassociation through HSS enabled unique person identification to our overall approach.

A real world empirical evaluation in a public train station ($N{=}420$) was presented. The results demonstrated self localisation robustly found accurate transforms, that our static image side-chain effectively enabled our approach to detect people physically interacting with environment, and that our HSS-based track association successfully resolves track ambiguities following merge/split events.

It was shown that our approach successfully detected and counted people without false positives in the complexities of a major Sydney public train station. Furthermore, these results show that our methods for robustness against sensor positioning and false negatives due to persons physically interacting with the environment, and for robust individual-specific-features based track association to resolve ambiguities contribute significantly to our overall approach and are real world viable. Furthermore, future work will aim at recognising instances in which the same person reappears after a period of time so that a unique persons count can also be generated along with the current count.

Acknowledgments

This work is supported by ARC Linkage Grant #LP110200708, the Centre for Autonomous Systems (CAS), RobotAssist (*www.robotassist.org*) and the Joint Centre for Robotics Research (JCRR) between UTS and Zhejiang University. Further information on our activities is available at *www.nkHRI.com.*

References

[Alempijevic *et al.*, 2013] A. Alempijevic, R. Fitch, and N. Kirchner. Bootstrapping navigation and path planning using human positional traces. In *Robotics and Automation (ICRA), 2013 IEEE Int. Conf. on*, pages 1242–1247, 2013.

[Browning *et al.*, 2006] Raymond C. Browning, Emily A. Baker, Jessica A. Herron, and Rodger Kram. Effects of obesity and sex on the energetic cost and preferred speed of walking. *Journal of Applied Physiology*, 100(2):390–398, 2006.

[Caraian and Kirchner, 2010] S Caraian and N Kirchner. Robust manipulability-centric object detection in time-of-flight camera point clouds. In *Proc. of the 2010 Australasian Conference on Robotics and Automation (ACRA)*, pages 1–8, Brisbane, Australia, 2010.

[Chen *et al.*, 2012] C.H. Chen, T.Y. Chen, D.J. Wang, and T.J. Chen. A cost-effective people-counter for a crowd of moving people based on two-stage segmentation. *Journal of Information Hiding and Multimedia Signal Processing*, 3(1):12–25, 2012.

[Garcia *et al.*, 2013] J. Garcia, A. Gardel, I. Bravo, J.L. Lazaro, M. Martinez, and D. Rodriguez. Directional people counter based on head tracking. *Industrial Electronics, IEEE Trans on*, 60(9):3991–4000, 2013.

[Hall, 1966] E.T. Hall. *The Hidden Dimension.* Doubleday, 1966.

[Hordern and Kirchner, 2010] D. Hordern and N. Kirchner. Robust and efficient people detection with 3-d range data using shape matching. In *Proceedings of the Australasian Conference on Robotics and Automation (ACRA)*, pages 1–9, Brisbane, Australia, 2010.

[Hsieh *et al.*, 2012] C.T. Hsieh, H.C. Wang, Y.K. Wu, L.C. Chang, and T.K. Kuo. A kinect-based people-flow counting system. In *Intelligent Signal Processing and Communications Systems (ISPACS), Int. Symp. on*, pages 146–150, 2012.

[Kirchner *et al.*, 2012] N. Kirchner, A. Alempijevic, and A. Virgona. Head-to-shoulder signature for person recognition. In *Robotics and Automation (ICRA), 2012 IEEE Int. Conf. on*, pages 1226–1231, 2012.

[Morton *et al.*, 2013] P. Morton, B. Douillard, and J.P. Underwood. Multi-sensor identity tracking with event graphs. In *ICRA*, pages 4742–4748. IEEE, 2013.

[Qiuyu *et al.*, 2010] Z. Qiuyu, T. Li, J. Yiping, and D. Wei-jun. A novel approach of counting people based on stereovision and dsp. In *Computer and Automation Eng., The 2nd Int. Conf. on*, volume 1, pages 81–84, 2010.

[Rabbani *et al.*, 2006] T. Rabbani, F. A. van den Heuvel, and G. Vosselmann. Segmentation of point clouds using smoothness constraint. In *IEVM06*, 2006.

[Zhao *et al.*, 2009] X. Zhao, E. Delleandrea, and L. Chen. A people counting system based on face detection and tracking in a video. In *Proc. of the 2009 Sixth IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pages 67–72, Washington, DC, USA, 2009.

[Zhu and Wong, 2013] L. Zhu and K.H. Wong. Human tracking and counting using the kinect range sensor based on adaboost and kalman filter. In *Advances in Visual Computing*, volume 8034 of *Lecture Notes in Computer Science*, pages 582–591. Springer Berlin Heidelberg, 2013.