Finding Tiny People: Long-Range Outdoor Sensing for Establishing Joint Attention in Human-Robot Interaction

by

Jake Bruce

B.C.S. Honours, Acadia University, 2013

Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Science

SIMON FRASER UNIVERSITY
Fall 2015

Approval

Name:	Jake Bruce
Degree:	Master of Science
Title:	*Finding Tiny People: Long-Range Outdoor Sensing for Establishing Joint Attention in Human-Robot Interaction*
Examining Committee:	**Dr. Anoop Sarkar** (chair) Associate Professor

Dr. Richard Vaughan
Associate Professor
School of Computing Science
Senior Supervisor

Dr. Greg Mori
Professor
School of Computing Science
Supervisor

Dr. Mark Drew
Professor
School of Computing Science
Examiner

Date Defended: 10 December 2015

Abstract

In this thesis, we present two novel methods for a robot to determine whether a distant human wants to interact. The first method finds periodic signals in camera streams and clusters them into potential human waving gestures. We demonstrated this on a ground robot by waving at the robot from long distances of up to 45 meters to attract its attention. The robot then approached to close range to confirm the gesture with other, more precise but distance-limited methods, such as human detection.

The second contribution is a robot behaviour for establishing joint attention with a human by exploiting trajectory signals. If a human is detected on an intercept course with the robot, the robot can vary its trajectory to probe the intent of the human. If the human corrects its trajectory to maintain the intercept, this can be considered a strong signal that the human wants to interact with the robot. We show that under modest assumptions, an arbitrary level of confidence in human intent can be achieved by iterating the behaviour.

In addition to contributions to human-robot interaction, we present a novel outdoor robot dataset captured at Simon Fraser University campus. The dataset consists of hundreds of gigabytes of trail data recorded using a ground-based robot equipped with six cameras and two laser scanners.

Keywords: mobile robotics, field robotics, human-robot interaction, gesture detection, spatial reasoning

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Just as capable computing systems have changed the way we manage information, capable outdoor robots have the potential to change the way we manage the physical world, from environmental monitoring to search-and-rescue. A crucial component of capable outdoor operation is the human factor: detecting and interacting with humans who require a robot's attention, or who want to modify a robot's behaviour. This thesis investigates two methods for establishing joint attention between humans and robots: a computer vision algorithm to detect periodic arm-waving gestures at long ranges, and a robot interaction behaviour that detects an intercept course with a human using arbitrary sensors such as the periodic gesture detector, and probes the human's intent before completing the rendezvous.

Detecting humans who are only 20 pixels tall in images is a serious challenge. As the first contribution, we propose and demonstrate an algorithm for detecting small clusters of periodically-oscillating regions of a consumer quality video stream. The system interprets these regions as strong signals of a potential human waving gesture, so the robot then approaches the signal to confirm the detection using more reliable human detection methods that require close range. We include results demonstrating this at ranges exceeding 45 meters using very poor VGA quality cameras.

The second contribution is an algorithm for determining whether a human is interested in interacting with the robot by analysing the relative trajectories of both entities. If a future intercept is detected, the robot varies its own trajectory to probe the intent of the human. If the human corrects its trajectory to maintain the intercept, then we consider this a strong signal of intent to interact with the robot. The robot can then complete the approach knowing the human is in fact interested. We demonstrate using GPS as the chosen sensor, although this method is not restricted to any particular modality. For example, intercepts could be detected using a visual technique such as the detector described as the first contribution of the thesis.

Chapter 2

Long-Range Gesture Detection

(a) View of a waving human at 35 meters (b) Human successfully approached

Figure 2.1: Human from robot's perspective: basic experiments

2.1 Waving Gestures

Consider the following situation: you are standing on a hill looking down into a crowd of people around a hundred meters away, attempting to find a friend. You are too far away to see any faces, and you don't know what color clothing she is wearing today. As you scan the scene, suddenly a repetitive motion catches your eye: you see your friend in the middle of the crowd, waving at you with both arms.

In a cluttered scene like the one described above, even the human visual system can fail to locate target objects until presented with a hint in the form of a color or a salient motion. In scenes that already contain a lot of motion, a repetitive action like a waving gesture can serve as a crucial clue to help direct the attention of the seeker. The person who wants to be found is injecting an unusual salient signal into the visual field of the seeker.

(a) View of a waving human at 35 meters (b) Human successfully approached

Figure 2.2: Human from robot's perspective: extended experiments

We are interested in methods for robots to cooperate with humans in large-scale outdoor environments. A useful component is to able to identify and approach humans over long distances where people can be as small as 20 pixels high, and against moving and cluttered backgrounds (see Figs. 2.1, 2.3). We demonstrate a human-robot interaction (HRI) system that uses consumer camera hardware to detect periodically oscillating image regions and identify candidate humans from long distances, after which the robot approaches the target for close-range interaction. Detections are based on periodic variation in pixel intensity over time, so we need make no assumptions about skin or clothing color, the texture of the background, or the precise scale of the human.

The contribution of this chapter is a long-range periodic gesture detection algorithm that reliably identifies periodic motions at distances of up to 35 meters using a low-resolution camera in which the human is roughly 20 pixels high. The HRI system in this chapter is the first to our knowledge that can locate and approach uninstrumented gesturing humans composed of so few pixels in indoor and outdoor environments, using only monocular camera sensing.

2.2 Related Work

Existing vision-based systems for uninstrumented HRI with low-resolution cameras require humans to be located less than ten meters from the robot to ensure they are composed of enough pixels to be identifiable. This is usually due to the use of face detection [1], [2], [3], skin detection [4], or model-based methods that require identification of particular body parts [5]. Action recognition methods [6] have been developed that operate at relatively long ranges (with humans as small as 30 pixels in height) but these assume a cropped figure-centric bounding box, which is difficult to extract when the human targets are very small or in front of a cluttered background. Discrimination between 24 distinct gestures has been

accomplished using frame-to-frame difference images [7] but once again the performance of this method degrades when the humans in the image are very small and dominated by noise. Optical flow-based techniques ([8], [9]) that rely on sparse features also tend to be unreliable for gestures at very long ranges, and computing real-time dense optical flow is not feasible on our robot due to limited computational resources.

The detection of periodic signals in image data has been under investigation for more than twenty years in the computer vision community, and a rich collection of approaches have been proposed. Some methods ([10], [11], [12]) track specific points or objects as they move through image space. Image alignment-based methods assume a figure-centric stabilized bounding box, and compute the self-similarity [13] or match points between periods of the oscillatory motion [14]. [15] makes use of aligned bounding boxes to compute a fast Fourier transform (FFT) of the pixels in the image over time, and fits the resulting frequency spectra to periodicity templates to discriminate oscillating pixels. We have experienced difficulty using feature tracking and image alignment methods on the scale of pixels, so we are interested in investigating other approaches.

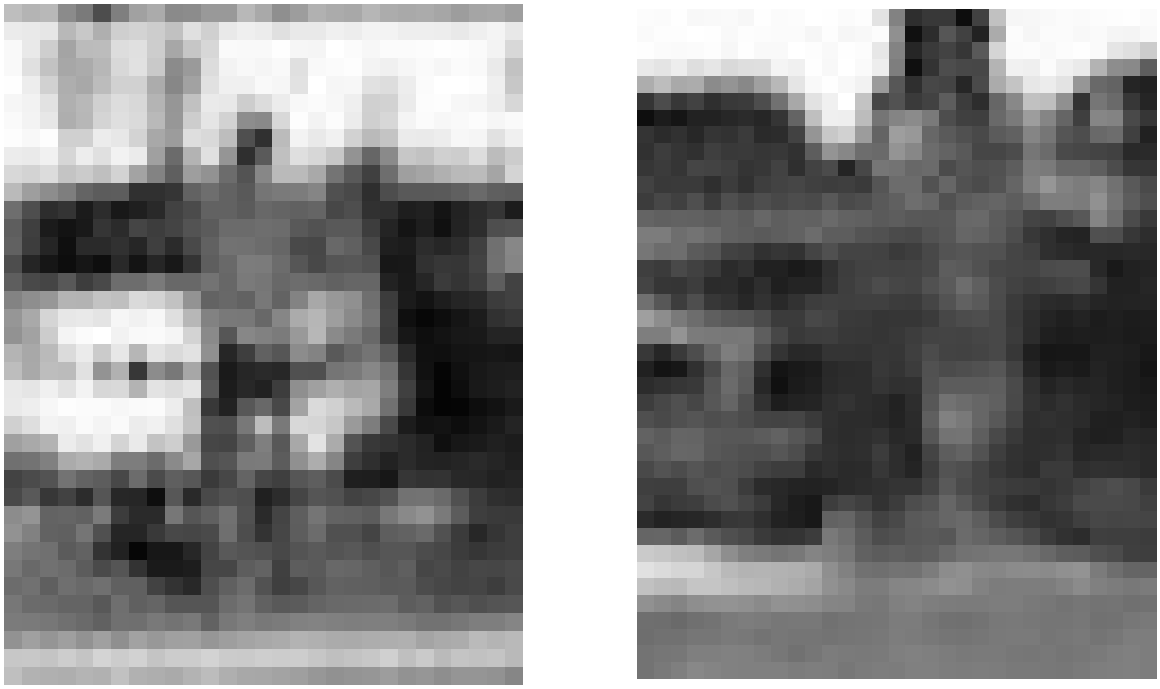
A pedestrian detection algorithm for infrared and color sensors has been proposed [16] that identifies human gaits using a periodicity metric called a *periodogram* [17], which is a quantitative measure of the degree to which a signal is periodic, based on the strength of the signal response at different frequencies. Periodic signal analysis has also been applied to long-range surveillance video [18] to identify walking pedestrians by analyzing the periodograms of blob trajectories, and by looking for an in-phase relationship between blob size and position. Segmentation-based techniques that rely on distinctive blobs become unreliable at long ranges, and currently a robot is more likely to include a visual camera than an infrared sensor due to relative cost.

Offline approaches have been developed to identify multiple periodic motions in video sequences by whole-video frequency and phase spectrum analysis [19], and to detect two-dimensional perspectives of oscillations in three-dimensional space using principles of affine invariance [20]. Affine invariance has the advantage of handling moving cameras, but these are not real-time methods.

Periodic signals in camera streams were exploited on the *Aqua* underwater robot to track and follow the oscillatory kicking motion of human divers at close range [21]. The FFT is performed on a time series of average pixel intensities in regions of the image, and significant peaks in the desired frequency are identified. A similar approach is described in [22] in which a support vector machine is trained to discriminate between gestures on the basis of frequency and phase spectra in an aggressively downsampled image.

We propose a real-time monocular vision method based on the *Aqua* robot system [21] combined with the periodicity metric from [12], which scores moving regions proportional to the relative strength of the fundamental frequency and its harmonics compared to the rest of the spectrum. The current chapter contributes a vision system that greatly improves the

maximum range of the state of the art for detecting waving human gestures by analyzing intensity changes for periodicity in very small regions of the image, and uses a flexible clustering algorithm to identify waving gestures in challenging environments using only low-resolution consumer camera equipment.



(a) Waving human in front of car

(b) Waving human in front of tree

Figure 2.3: Successful difficult detections

2.3 Periodic Motion

The proposed algorithm for periodicity detection can be described in three stages:

A) constructing a time series of average pixel intensity on a per-region basis

B) identification of periodic signals in the desired frequency range for human waving

C) clustering periodic regions into large-scale bounding boxes for output

This section describes each stage in detail.

2.3.1 Intensity Time Series

To construct a set of time series buffers to check for periodicity, we divide the image into a set of regions of interest and compute the average grayscale intensity of the pixels in each region over time. Since we are interested in detecting people on the order of 20 pixels in height (see Fig. 2.3), we use regions 10 pixels on a side, which we overlap by half along each axis. Using smaller regions increases the range of the detector by allowing it to detect smaller motions, but also increases the computational demand due to greater region count.

Our system requires that the robot be stationary in order for these image regions to remain in place as time goes on. Assuming the robot is stationary simplifies data association between frames; in parallel work [23] we have removed this constraint.

A weighted average grayscale intensity of the pixels in each region is computed using a Gaussian kernel centered on the middle of the region. This non-uniform weighting reduces edge effects between regions, and ensures that a pixel that only moves inside one region still produces variation in its box's average intensity. We use a symmetric Gaussian kernel with $\sigma = 5$. This step is roughly equivalent to downsampling the image as in [22].

These weighted averages are stored in a circular buffer with a 2 second time horizon for each region: a typical series is shown in Fig. 2.5a. The length of the temporal window should be chosen to include at least two periods of the gesture in order for the periodicity to be clearly present in the frequency spectrum. Increasing the length of this window increases robustness to false positives, but slows the response time of the detector.

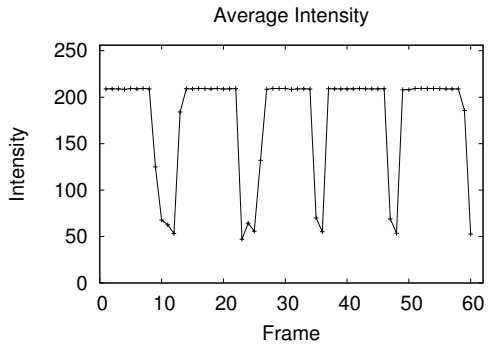
2.3.2 Evaluating Periodicity

A periodic signal is composed of a signal oscillating at a fundamental frequency plus its harmonics. In this application, periodic signals are embedded in noisy time series data, so the system discriminates time series that contain sufficiently strong periodic signals from those that do not, on the basis of the frequency spectrum of the signal (see Fig. 2.5b). To make this distinction we use a metric proposed in [12] in which the periodicity P_F of a signal with power spectrum F and highest amplitude frequency w is given by:

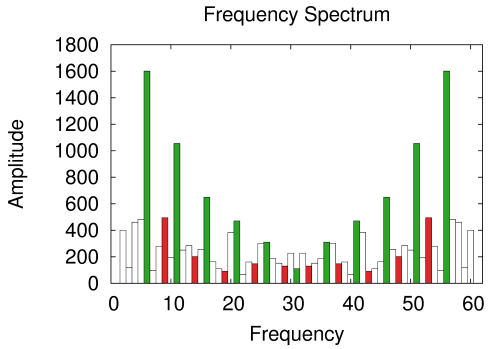
$$P_F = \frac{\sum_i F_{iw} - \sum_i F_{iw+w/2}}{\sum_i F_{iw} + \sum_i F_{iw+w/2}} \tag{2.1}$$

This quantity is a normalized difference between the sum of the power spectrum values at the highest amplitude frequency and harmonics, and the sum of the values at the frequencies halfway between. This yields a score indicating the relative strength of the frequency w and harmonics compared to the rest of the spectrum. Signals with P_F near 1 are highly periodic, and P_F values close to 0 describe signals with little to no regular oscillation. Fig. 2.5b shows a frequency spectrum with DC component removed. Green frequency components indicate the fundamental frequency and its harmonics (F_{iw}), while red indicates the components

Figure 2.4: Example detected periodic gesture



(a) Intensity signal from a waving gesture



(b) Corresponding frequency spectrum

Figure 2.5: Time series and frequency spectrum of periodic signal

halfway between ($F_{iw+w/2}$). The red components are chosen to be as far as possible from the green components to avoid aliasing effects due to the finite sampling frequency. The periodicity P_F of this signal is 0.61.

We consider a signal to represent a potential gesture if $P_F > 0.45$ and w is between 1Hz and 3Hz, as humans tend to wave at approximately this rate. In addition, we only consider signals with fundamental amplitude $F_w > 30$, to avoid false positives due to minor lighting fluctuations and compression artifacts. These criteria define the sensitivity of the system to noisy periodic signals, and the frequency range of the target gesture. The values of these parameters were chosen based on informal experiments, but in our experience this configuration successfully detects periodicity for a variety of gestures and in domains other than pixel intensity, such as pixel locations in image space.

We evaluate the periodicity of each time series buffer every N frames, which can be chosen based on available processing power. We choose $N = 15$ on our system for an operating frequency of 2Hz. With more computational power the algorithm can be run at

a faster rate, which helps eliminate false positives caused by transient apparent periodicity. We filter the output, requiring a hit-to-miss rate of at least 3 : 1 in a region over a 2 second time window before accepting it as a potential stationary waving human. Increasing the length of this window or the required hit-to-miss rate improves robustness to transient periodicity, but slows the response time of the detector.

2.3.3 Clustering

Given a collection of 10×10-pixel regions flagged as positive for periodic motion, the system forms large-scale bounding boxes (see Fig. 2.4) to identify multiple sources of motion if present. We use the scikit-learn implementation [24] of the DBSCAN algorithm [25] which clusters unlabeled data by forming connected subgraphs and makes no assumptions about the number of clusters.

DBSCAN requires two parameters: the maximum distance ϵ between connected data points and the minimum size δ of a cluster. Any detections more than ϵ pixels away from a connected group of at least δ other detections are considered outliers and do not affect the output of the detector. We use $\epsilon = 45$ pixels and $\delta = 3$, chosen to form sensible clusters at both short and long ranges. This parameter depends on the expected size of the subjects in image space, and works well for our application of clustering human gestures 20 pixels and larger in size.

2.4 Gesture Approach Behaviour

A detector that only functions while the robot is stationary presents a challenge for defining behaviour. We cannot simply servo to the detection, as camera movement can produce apparent periodicity in non-periodic scenes. The position of the stationary detection along the horizontal axis of the image is sufficient to drive our robot accurately in the direction of the target, but how far the robot should travel is not obvious. The goal of the system is simply to get within range of more discriminative sensors such as face, torso, or human detectors, so the robot drives in the direction of the target for a distance of 10 meters, at which point the robot stops and makes another stationary scan to correct for angle error during detection and approach.

In the case of multiple detections, we choose a potential human to approach based on the persistence of the gesture. Walking pedestrians, vehicle traffic, waving flags and trees can all cause apparent periodicity; however, periodic motion from a stationary human tends to persist, while false positives tend to come and go. Our robot behaviour waits until exactly one detection is reported before approaching. This also helps at close range, where the two hands of the human can be clustered as two separate periodic motions.

Chapter 3

Evaluating Periodicity Detector

Figure 3.1: Experimental setting: basic experiments

3.1 Basic Experiments

We first tested the system under relatively easily conditions—uncluttered and open outdoor terrain—to estimate effective range and detection rates under ideal conditions. Experiments were performed using a Husky A200 ground-based robot built by Clearpath Robotics. Excluding emergency-stop behaviour, the only active sensor for these experiments is a consumer 640×480 resolution monocular Kinect camera mounted on the front of the robot, providing color video at 30 frames per second. The robot includes an onboard computer with 8GB of RAM and a dual-core Intel Core i5 2.4GHz (2012 laptop-class) processor.

We find that with 10×10 regions and a 640×480-pixel image, our system can consistently identify periodic waving gestures by humans wearing a variety of colors and on different backgrounds, out to a distance of approximately 35 meters. Beyond this distance detections become unreliable in the absence of strong contrast between subject and background, although in ideal cases with high contrast the system successfully identifies periodic motions at distances of up to 60 meters. Consistent detection at longer distances can be achieved with the use of regions smaller than 10 pixels on a side and/or image resolutions greater than 640×480, but the increased computation required is not feasible on our robot at this time.

Fig. 3.1 shows a panoramic view of the experimental setting: an outdoor area on Simon Fraser University campus with pedestrian and vehicle traffic, trees, and flags in

Figure 3.2: Husky A200 robot

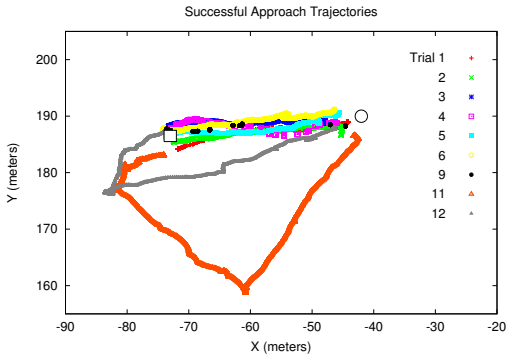
view. Each trial begins with the robot in a particular orientation, located 35 meters from an uninstrumented human. The human performs a two-arm waving gesture as shown in Fig. 2.1 until the robot stops within 10 meters, which is considered a successful approach. If the robot does not arrive within 180 seconds, the trial is considered a failure.

We recorded two approaches per orientation for a total of 12 trials, with the following robot orientations in degrees relative to the human target: $\{-45, 0, 45, -135, 180, 135\}$. With the field of view of our robot, the human subject is initially visible in the first three orientations only. Results are shown in Table 3.1, and approach trajectories are shown in Fig. 3.3.

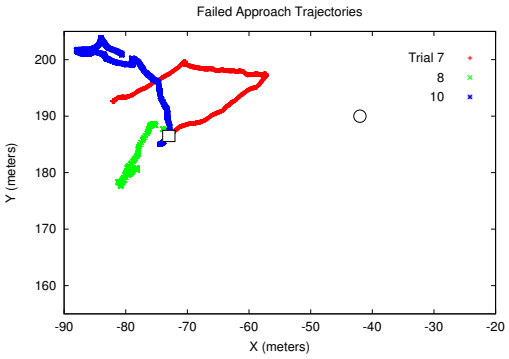
Nine out of twelve approaches resulted in the robot arriving within 10 meters of the human in less than 180 seconds, for a total success rate of 75 percent. Given more time, the failed approaches may have eventually arrived at the target, but occasionally a distractor such as a tree or flag caused the robot to investigate in the opposite direction from the human. This resulted in the robot exceeding the maximum range of the detector and being unable to return except by chance. As shown in Fig. 3.3a, on two occasions the robot made significant detours to investigate waving trees, and in both cases the trees were correctly rejected at close range followed by the robot successfully approaching the human.

Trial		Orientation (°)	Time (s)	Success/Fail
1	↑	0	75	S
2			60	S
3	↖	45	75	S
4			70	S
5	↗	-45	68	S
6			81	S
7	↘	-135	—	F
8			—	F
9	↙	135	69	S
10			—	F
11	↓	180	154	S
12			93	S

Table 3.1: Results: basic experiments



(a) Approach trajectories that reached the human. In trials 11 and 12, the robot investigated and rejected trees waving in the wind by approaching and using HOG human detection at close range.



(b) Approach trajectories that failed to reach the human. Failed trials often approached a false positive detection such as a tree, ending up too far from the human to recover.

Figure 3.3: Trajectories: basic experiments

3.2 Extended Experiments

Once the effective range and performance of the detector had been estimated in the basic experiments, we designed extended experiments to test the system under more realistic and challenging conditions: a cluttered and moving background, with moving human and non-human distractors.

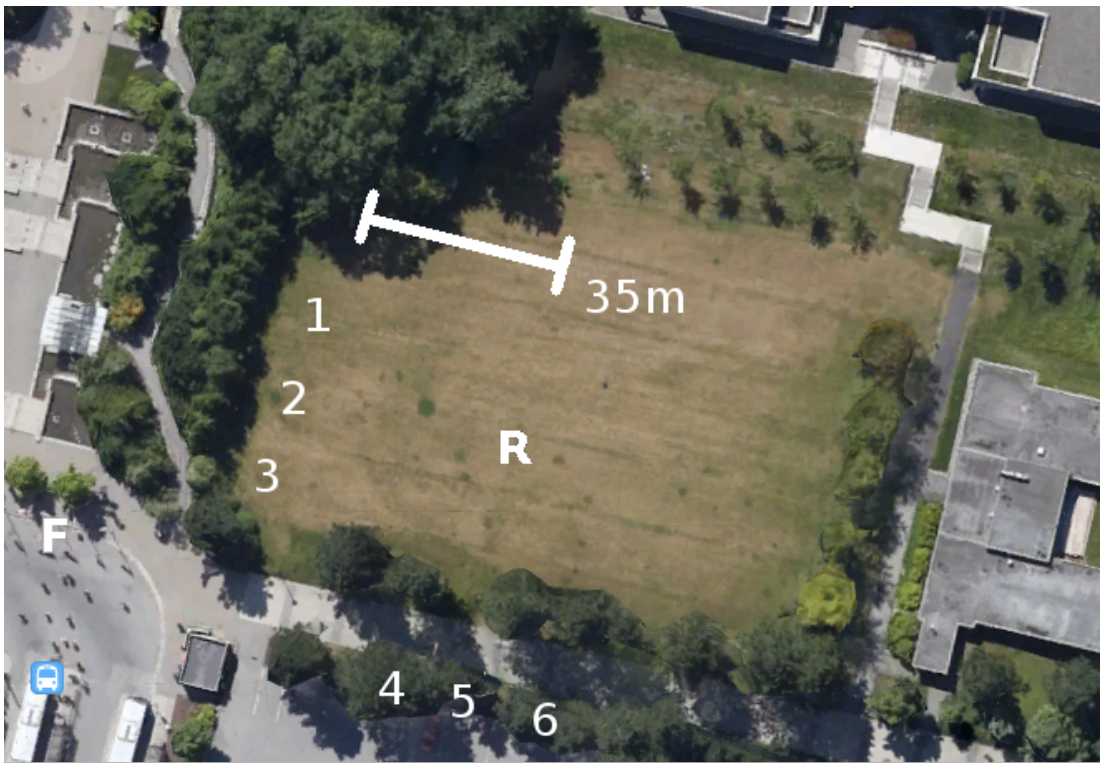


Figure 3.4: Experimental setting: extended experiments

Fig. 3.4 shows an aerial view of the experimental setting: an outdoor area on Simon Fraser University campus with frequent natural pedestrian and vehicle traffic, waving trees and flags in view (at location F in the image). We performed two sets of experiments: a distance investigation in which detection rates are recorded at different ranges between the robot and four subjects in three locations; and an approach investigation in which the robot attempts to approach to within 3 meters, using the same subjects and locations as the first scenario. Both experiments were repeated against two different backgrounds, shown in Fig. 3.5, and all trials were performed using the same values for all parameters: the values reported in this thesis. Importantly, the human subjects and locations 1 through 6 were chosen arbitrarily without testing the detector against these subject/background combinations, in order not to bias test results. In both distance and approach tests, the human subject performed a two-arm waving gesture as shown in Fig. 2.4.

(a) locations 1, 2, 3

(b) locations 4, 5, 6

Figure 3.5: Robot's perspective: extended experiments

	Locations 1,2,3			Locations 4,5,6		
Distance	Success	Time	Rate	Success	Time	Rate
25m	12/12	8.3s	100%	11/12	15.2s	92%
30m	12/12	9.5s	100%	11/12	14.5s	92%
35m	11/12	11.9s	92%	9/12	17.3s	75%
40m	10/12	11.7s	83%	9/12	18.0s	75%
45m	9/12	15.9s	75%	4/12	24.8s	42%

Table 3.2: Results: distance experiments

3.2.1 Distance Experiments

In order to test the effective range of our detector on a 640x480-pixel camera against varying backgrounds, we tested whether the system could detect a human waving at distances of 25, 30, 35, 40, and 45 meters. We evaluated the system at each distance using a stationary robot with four different people standing in three different locations for a total of 12 trials against each of the two background environments. We considered a trial successful as soon as the detector reported a bounding box containing the gesturing human. If the system did not detect the gesture within 60 seconds, the trial was considered a failure. Results are reported in Table 3.2.

To analyze the statistical significance of the distance results, we compare against an imaginary detector that chooses a pixel uniformly at random as the center of the detection, and we consider it a successful detection if the chosen pixel is inside the target bounding box. For an SxS-pixel bounding box in a 640x480-pixel image, the probability of success per frame is the ratio of the area of the box to the area of the image.

If we imagine that we match our real trials by running this random detector once every $N = 15$ frames of the video sequences from the distance experiments, the probability that the random detector would successfully identify the waving human over a 60 second window in one of these trials is given by the complement of the probability of not finding the human:

$$p_{trial} = 1 - (1 - \frac{S^2}{640 * 480})^{120} \tag{3.1}$$

We test the significance of our results against the binomial distribution using $n = 12$ and $p = p_{trial}$ for the size of the human at each distance. We reject with 99% confidence the null hypothesis that our detector is no better than random in every case, except for the 45 meter case with success rate of 4/12, for which we can only say with 85% confidence that our system is better than the imaginary random detector.

Location	1	2	3	4	5	6
Subject 1	✓	✓	✓	✓	✗	✓
Subject 2	✓	✓	✓	✓	✓	✗
Subject 3	✓	✓	✓	✓	✗	✓
Subject 4	✓	✓	✓	✓	✓	✗
Success rate:	100%			66.6%		
Overall rate:	83.3%					

Table 3.3: Results: extended experiments

3.2.2 Approach Experiments

Given an estimate of the success rate of the detector at different ranges, we chose a distance for approach experiments that was likely to find the person but also demonstrate the value of the system in approaching from long range. Our approach experiments were performed with the robot starting 35 meters from the subject, facing toward the middle location in each setting. The robot scanned for waving gestures and drove toward the first detection for 10 meters before beginning another scan.

We defined an approach as successful once the robot drives to within 3 meters and the person was fully visible in the camera image. The scan-approach behaviour continued until the robot stopped in front of an obstacle, which would be the human target if the approach succeeded. If the robot did not arrive within 180 seconds, the trial was considered a failure. Results are reported in Table 3.3.

In addition to the waving subject, both environments contained natural and artificial distractors and frequent occlusions of the target. Lighting conditions varied gradually throughout the duration of trials in both environments as blue skies transitioned to clouds. For locations 1, 2 and 3, we planted two stationary humans and four moving humans instructed to wander randomly around the area at varying distances to the robot, and often in front of the subject. Distractors also visible in this environment include intermittent bus traffic, waving tree foliage and flags blowing in the wind (2 to 7 kilometers per hour).

For locations 4, 5 and 6, the waving subject was located on the opposite side of a busy pedestrian walkway against a background of parked cars. Most of the distractors in this environment were natural, non-informed human pedestrians who occluded the subject at short intervals, typically between one and five seconds. At times when natural pedestrian traffic was thin, we injected informed humans into the walkway to maintain a roughly consistent occlusion and distraction rate for all trials. In addition to human distractors, this environment contained occasional vehicle traffic and trees moving in the wind, at similar wind speeds as the first set.

3.3 Discussion

The distance evaluation shows that our system works consistently out to longer ranges than any known real-time method for locating waving humans. It also indicates the effect of contrast to background, and the effect of occlusions and distractors (both natural and human), as the maximum range drops noticeably in locations 4, 5 and 6, and the average time increased considerably over the first environment. This is due to both the challenging background and the tendency of passing pedestrians to obscure the subject, requiring more samples to identify the periodic gesture.

Light and shadow is also a factor, as locations 5 and 6 were partly in shadow which reduced the contrast of the human against the background and resulted in the failure of the approach involving subject 2, who was standing in shadow in front of a dark blue car wearing a blue jacket. This rendered the subject essentially invisible in both grayscale and color images, so the robot did not move at all during this trial.

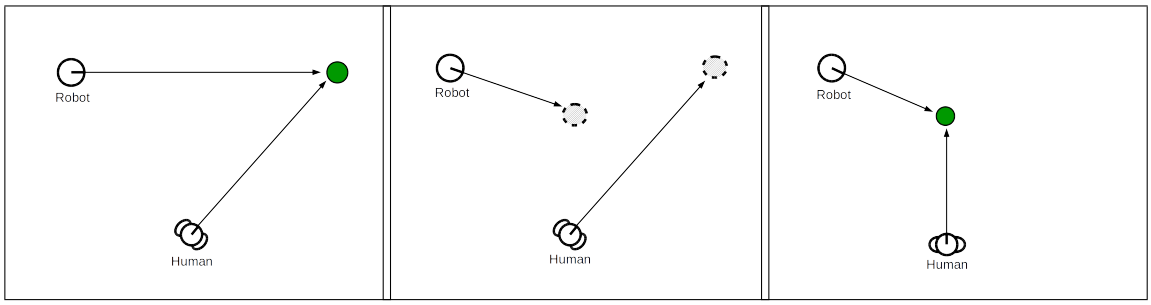
Shadows can also benefit the detector under the right circumstances, as we observed during periods of bright sunlight where the shadow (oscillating along with its human source) exaggerated the size of the apparent periodic motion and caused it to be detected at further ranges than would be likely under cloudy weather.

Several failures were due to the conversion of images from color to grayscale. In the approach failures involving subjects 1, 3, and 4, several pedestrians wearing different colored clothing walked by at a roughly constant interval, which can appear periodic in grayscale and caused the robot to drive in the wrong direction. Due to the many-to-one mapping of the grayscale manifold, colors which are visibly distinct in color space are often compressed down to the same or similar grayscale values, as in the case of the red plaid worn by the subject on the right in Fig. 2.3 against the brown of the tree. This can result in subjects becoming effectively invisible in grayscale even when they are clearly visible in color images, and can make aperiodic color sequences appear periodic in grayscale.

One approach to reducing this effect involves running periodicity checks on all three color channels, detecting regions as positive for periodic motion if any of the channels are moving periodically. Such a system should also reject regions containing aperiodic but non-stationary signals in any of their color channels, as in the example case of $black \rightarrow green \rightarrow black \rightarrow yellow \rightarrow black \rightarrow brown \rightarrow black$, which should not trigger a detection: the green channel is periodic, but the other channels are aperiodic and non-stationary. This increased sensitivity can expose the system to false negatives which would not have been missed in grayscale though, so we leave this extension for future work.

Chapter 4

Determining User Intent



(a) Robot and human on an intercept course

(b) Robot varies its trajectory to remove the intercept

(c) Interested human adjusts to maintain an intercept

Figure 4.1: Trajectory behaviour sequence

4.1 Trajectory Signals

Many human-robot interaction (HRI) systems provide feedback to the human user in the form of auditory or visual cues [26]. Users must often be instructed in how to interpret the cues or behaviour of the robot before useful interaction can be performed, and we have found that these cues tend to be most effective at close ranges. We are interested in developing HRI interfaces and behaviours that are natural and intuitive, require minimal special equipment and training, and work at long ranges.

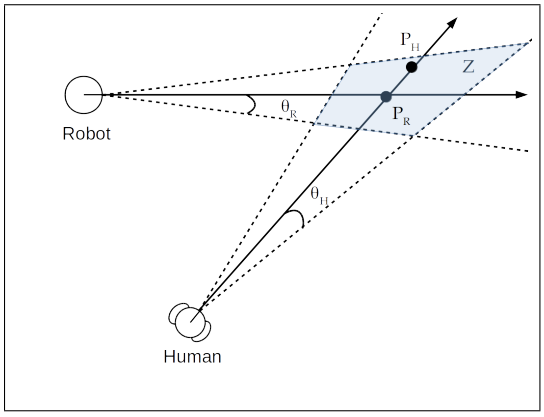
An established approach to achieving robust and natural HRI involves a dialogue between human and robot, where the two agents dynamically converge to agreement about the task at run-time [27]. In this chapter, we propose a behaviour pattern for co-operative human-robot rendezvous that uses motion trajectories, a feature common to all mobile agents, as an interaction signal to determine whether a distant human is interested in interacting with the robot [28].

4.1.1 Trajectory Behaviour

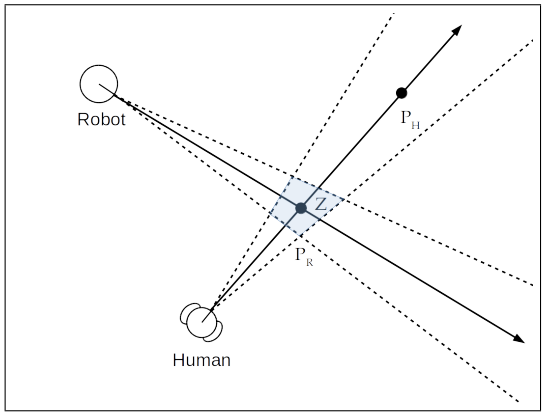
Our co-operative rendezvous behaviour requires only that the approximate geometry, relative position and velocity of the human and robot are known, and does not assume any specific sensor modality. The behaviour sequence (Fig. 4.1) begins during a default behaviour (random wandering in our case) by continuously computing whether the trajectory of the human will intercept the robot before some maximum time. If the human is on an intercept course with the robot, the robot then tests this coincidence by varying its trajectory until it has removed the intercept. If the human corrects its own trajectory to recreate an intercept, we consider this a strong signal that the human intends to interact, and the robot continues along the current trajectory until intercepting. At this point, close-range interaction techniques such as gesture or speech recognition can be used ([2], [29]) with confidence that the user intends to interact with the robot.

4.1.2 Intercept Detection

In order to detect future intercepts, we use simple spatio-temporal geometric reasoning on a 2D plane, as shown in Fig. 4.2. Since we know the heading of the robot and human to within uncertainty of θ_R and θ_H respectively, we can create trajectory cones representing sets of potential future locations given the current orientation. Trajectory cones have the desirable property of widening with distance, which represents intercept uncertainty growing as range increases.



(a) Detecting intercepts from position and velocity. Z is the shaded intercept zone between trajectory cones. P_R and P_H are the expected locations of the robot and human at time t. Here, P_H is within Z



(b) Detecting a non-intercept trajectory: P_H is outside of Z

Figure 4.2: Detecting intercepts from position and velocity

We define the spatial region of intersection between the two cones as a potential intercept zone Z, and compute the time t for the robot to reach the center of mass of this region at its current velocity. As in [30], we model the geometry of the robot and human by planar discs, and report an intercept only if at time t the human is expected to be located inside Z or closer to P_R than the combined radius of the two agents, in the case that Z is small.

4.2 Confidence

We have claimed that correcting one's trajectory to maintain an intercept is a strong signal of intent to interact. The strength of the signal depends on the characteristics of the sensor used to detect the intercept, as well as behaviour of the environment and the other agents in it. We can extend the system so described to cycle the behaviour an arbitrary number of times to achieve a desired level of confidence in our decision. Assuming that a corrected intercept means intent to interact with probability p, then we can express our confidence c that a particular human wants to interact:

$$c = 1 - (1 - p)^n \tag{4.1}$$

Where n is the number of behaviour iterations in which the user corrects its trajectory to maintain the intercept. Equivalently, we can express the number of times a robot would need to iterate the behaviour sequence to reach an arbitrary confidence:

$$n = \lceil (\frac{\log (1 - c)}{\log (1 - p)}) \rceil \tag{4.2}$$

4.3 Demonstration

We have demonstrated that trajectory signalling provides a basis for successful discrimination and rendezvous with interested humans in an outdoor environment, using filtered GPS signals from consumer mobile phones to determine the position and velocity of both the robot and human. In recorded video, we first demonstrate that the robot will ignore humans who are not attempting an intercept trajectory, and then show that interested humans attempt an intercept trajectory, successfully signal to the robot and complete the rendezvous. Informally, the system performs as described and feels natural to use. The video has been provided at `www.youtube.com/watch?v=_gV8cmUsIhA` showing the system in operation.

Chapter 5

SFU Mountain Dataset

(a) Trans-Canada Trail (b) Powerline Trail (c) Jim's Jungle Trail

Figure 5.1: SFU Mountain Dataset trails

5.1 Overview

In addition to the HRI work described above, we have also produced an open access dataset of outdoor robot navigation [31]. Although unrelated to HRI, we will describe the dataset here, as an additional contribution to outdoor robotics.

The SFU Mountain Dataset is a novel long-term dataset of semi-structured woodland terrain under varying lighting and weather conditions and with changing vegetation, infrastructure, and pedestrian traffic. This dataset is intended to aid the development of field robotics algorithms for long-term deployment in challenging outdoor environments. It includes more than 8 hours of trail navigation, with more available in the future as the environment changes. The data consist of readings from calibrated and synchronized sensors operating at 5 Hz to 50 Hz in the form of color stereo and grayscale monocular camera images, vertical and push-broom laser scans, GPS locations, wheel odometry, inertial measurements, and barometric pressure values. Each traversal covers approximately 4 km across three diverse woodland trail environments, and we have recorded under four different lighting and weather conditions to date: *dry*; *wet*; *dusk*; *night*. We also provide 383

hand-matched location correspondences between traversals as ground-truth for benchmarking place recognition and mapping algorithms. This chapter describes the configuration of the vehicle, the trail environments covered, and the format of the data we provide.

The SFU Mountain Dataset was recorded on a mobile ground-based robot driving from the summit to the base of Burnaby Mountain, British Columbia, Canada, covering an altitude change of nearly 300 m (Figure 5.2). Sensors include color stereo cameras, monocular grayscale cameras, vertical and push-broom scanning laser rangefinders, GPS, wheel encoders, an inertial measurement unit, and a barometric pressure sensor.

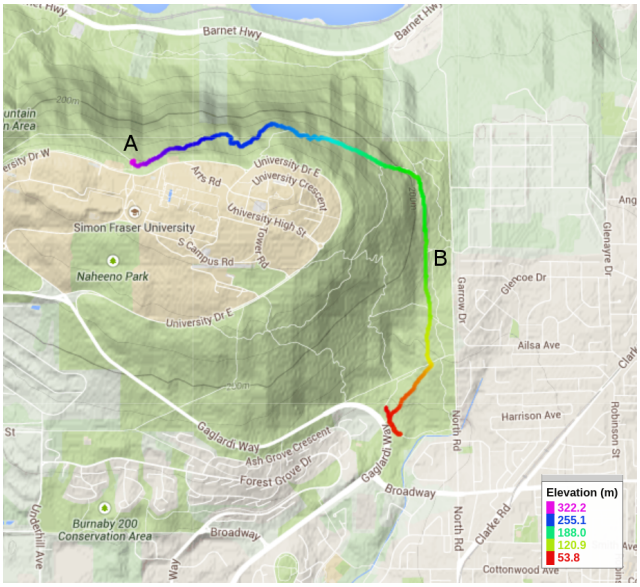


Figure 5.2: GPS map of SFU Mountain Dataset trails

The main purpose of the dataset is to provide comprehensive coverage of several types of semi-structured woodland trails under changing conditions (i.e. lighting, weather, vegetation, infrastructure, and pedestrians) in a highly self-similar natural environment. These data differ from most existing offerings such as the KITTI dataset [32], which covers structured urban environments targeted toward developing autonomous car technology. In contrast, we traverse challenging semi-structured woodland trails, resulting in data useful for evaluating place recognition and mapping algorithms (*i.e.* [33], [34]) across changing conditions in natural terrain.

The data, approximately 150 GB in size at this time, can be downloaded from `http://autonomylab.org/sfu-mountain-dataset`. We provide sensor data exported as JPEG images and CSV text files, and also the ROS bag files that were recorded directly from the robot.

5.2 Robot Setup

The configuration of our recording platform is illustrated in Figure 5.4:

- 2 × PointGray Firefly color cameras facing forward in stereo configuration with approximately 90° field of view (FMVU-03M2C-CS), 752 × 480 pixels, 1/3" Aptina MT9V022 CMOS, global shutter, 30 Hz
- 4 × PointGray Firefly monochrome cameras facing port, starboard, rear, and upward with approximately 90° field of view (FMVU-03M2M-CS), 640 × 480 pixels, 1/3" Aptina MT9V022 CMOS, global shutter, 30 Hz
- 1 × SICK LMS111 scanning laser rangefinder with 270° field of view in 0.5° increments, mounted with 180° roll and angled toward the ground in "push-broom" style approximately 20° to the horizontal, 18m range, 50 Hz
- 1 × SICK LMS200 scanning laser rangefinder with 180° field of view in 1° increments, sweeping a vertical plane normal to the x-axis of the robot, 8 m range, 10 Hz
- 1 × Garmin 18x GPS receiver with 15 m accuracy at 95 % confidence, 5 Hz
- 1 × UM6 inertial measurement unit providing orientation with 2° pitch and roll accuracy and 5° yaw accuracy, angular velocity and linear acceleration, 50 Hz
- 1 × Wheel encoders providing linear and angular velocity at 10 Hz
- 1 × Barometric pressure sensor from LG Nexus mobile phone in Pa at 30 Hz
- 4 × Titan 54 W off-road LED lights (ORBT9-54WD-FL) with brightness of 3780 lm, 5000 K color temperature and 60° beam angle, night sessions only

All cameras have exposure set to automatic, resulting in large shifts in effective brightness and in the amount of motion blur, which is significant in lower lighting. Color cameras are Bayer filtered, resulting in less detailed images than those from the grayscale cameras. During the night session, lights are mounted on the base plate below the two stereo cameras pointing outward at approximately 10° to the cameras' optical axes, as well as one light mounted above each side camera pointing in the port or starboard direction, covered with white tissue paper to improve light diffusion.

The robot was driven at its maximum speed of 1 m/s for most of the dataset, except for rough sections of Jim's Jungle.

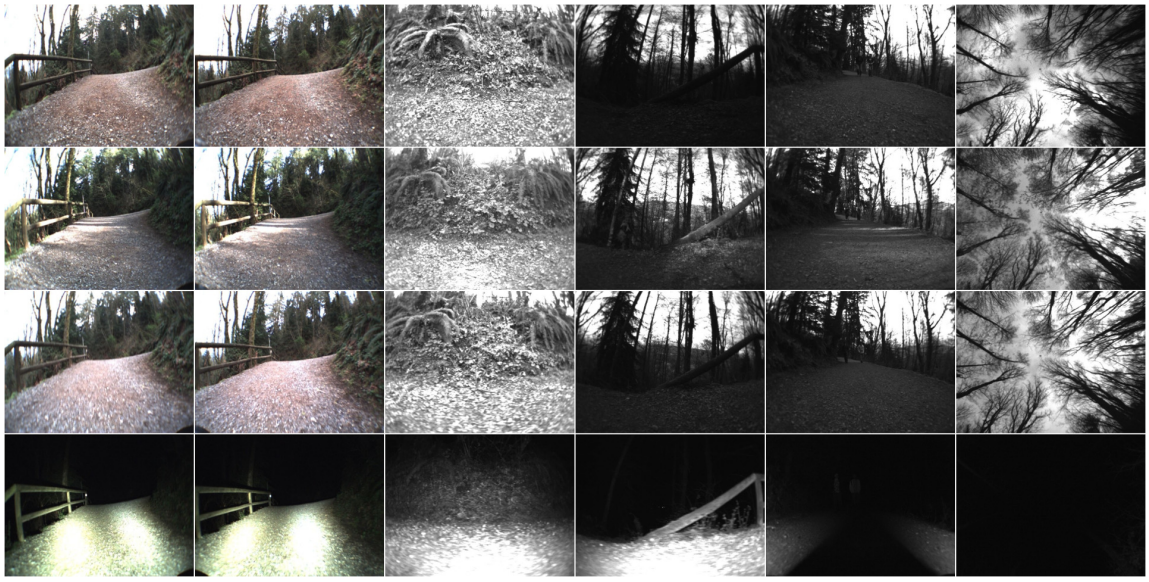


Figure 5.3: Sample place match from SFU Mountain Dataset

5.3 Dataset

The dataset covers the traversal of three trails from the summit to the base of Burnaby Mountain, with a battery swap break approximately halfway through. We call the first and second halves of the data *part A* and *part B*, the start locations of which are marked in Figure 5.2. Histograms of sensor readings are shown in Figure 5.6 to summarize and compare the statistics of the two parts.

Part A includes the Trans-Canada Trail and approximately half of the Powerline trail, while part B consists of the rest of the Powerline trail and several hundred meters of Jim's Jungle trail. Recordings were made in four environmental conditions, which we refer to as *dry*, *wet*, *dusk*, and *night*.

5.3.1 Trail Environments

Trans-Canada Trail — densely-forested mountainside terrain with a gray gravel path approximately 3 m wide. The starboard side of the path faces up the slope of the mountain, and is mostly dirt and small vegetation such as ferns and moss, with occasional tree trunks. The port side of the path faces down the slope, looking out on small vegetation and dense tall trees, with water and mountains in the distance. This section of the dataset covers an altitude change of approximately 125 m. The Trans-Canada Trail section consists of challenging and self-similar terrain, but distinctive natural and artificial landmarks are common.

Powerline trail — cleared woodland terrain on a gray and brown gravel path averaging 3 m wide, with low bushes and distant trees on both sides. This trail section includes powerlines, wooden power poles, and steel powerline towers. Most of this segment is oriented along the North-South cardinal axis. The Powerline trail is highly self-similar with few unique landmarks, and covers an altitude change of approximately 150 m.

Jim's Jungle — a section of trail at the base of Burnaby Mountain with dense tree cover and a narrow brown dirt path approximately 1 m wide. This segment has frequent turns, little altitude change, and an uneven trail surface that causes sharp orientation changes and occasional wheel slippage. On sunny days, shadows and bright patches are more common and more severe than in the other sections due to the dense canopy.

5.3.2 Conditions

- *dry*—recorded April 19, 2015 on a sunny day in good weather, with strong shadows and occasional severe light interference in the camera images.
- *wet*—recorded March 24, 2015 on a rainy day with overcast skies; shadows mild or nonexistent. The second half of part A contains a stop to attach an umbrella, which protects the vehicle from the rain and obscures the upward camera. The rain configuration of the vehicle is shown in Figure 5.5a.
- *dusk*—recorded April 2, 2015 on a dry overcast day just before sunset. The environment has an ambient brightness of approximately 400 lx at the beginning of part B, declining to nearly 30 lx by the bottom of the Powerline trail, and is almost zero lux in Jim's Jungle.
- *night*—recorded April 20, 2015 in dry weather, long after sunset. Bright off-road LED lights were mounted on the front and sides of the robot for this session to illuminate objects near to the robot, with brightness dropping off quickly beyond a distance of several meters. Figure 5.5b shows the vehicle on the Powerline trail at night.

5.3.3 Ground Truth Locations

In addition to the sensor data, we provide 383 ground-truth location matches between the four sessions: 237 from part A and 146 from part B. These are hand-aligned locations separated by approximately 10 m according to GPS readings, and provide a set of correspondences between the sessions by timestamp and sets of matching camera images. Envisioned uses include evaluating place recognition algorithms on known place matches, or for establishing known correspondences between localization and/or mapping systems over the different sessions. Figure 5.3 shows a single location from the Trans-Canada Trail recorded by each camera across all four conditions.

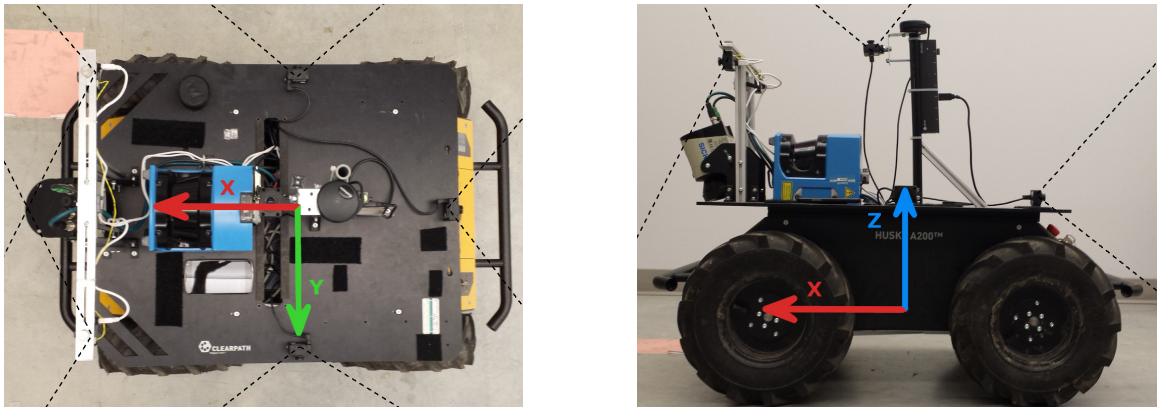


Figure 5.4: SFU Mountain Dataset coordinate frames

(a) Wet mode

(b) Night mode

Figure 5.5: Alternate Husky configurations

5.3.4 Sensor Calibration

We provide spatial transformations for each sensor in the form of a vector in $\mathbf{R}^6$, which represents x, y, z translation in meters and roll, pitch, yaw in radians with respect to the robot's origin as shown in Figure 3.2. Orientation is applied in the order of roll, then pitch, then yaw with respect to the fixed axes of the robot's coordinate frame. For cameras, we also provide intrinsic calibration in the form of a 3×3 camera matrix and 5-parameter plumb bob distortion model, in the form used by OpenCV.

We have synchronized sensor timestamps by measuring the rate of change of each sensor when the robot starts moving, and aligning the spikes representing this common event to the same time by a fixed offset. The timestamps of the bag files and in the CSV files already incorporate this offset, which is given on the web page of the dataset for reference. The only sensors for which we cannot synchronize timestamps are the GPS and pressure sensors, which are fortunately also the least time-sensitive: neither pressure nor GPS location are as precise as the other sensors. Timestamps in the CSV files are given in nanoseconds since January 1, 1970.

5.3.5 Data Format

Data is available in the form of JPEG image files, CSV text files and ROS bag files recorded directly from the vehicle. Parts A and B of the trail sequences are available as separate gzipped archive files `<sensor>-<session>-<part>.tgz` and bag files `<session>-<part>.bag`. The first line of each CSV file is a comma-separated header labeling each comma-separated field in the file. Images are bundled by camera and are named by their timestamp.

In general, data are given in raw form except for the timestamp offset. However, we provide GPS locations in both (lat, long, alt) and (x, y, z) forms, with the latter given in meters in ENU (East-North-Up) coordinates with origin at the location of the first GPS location.

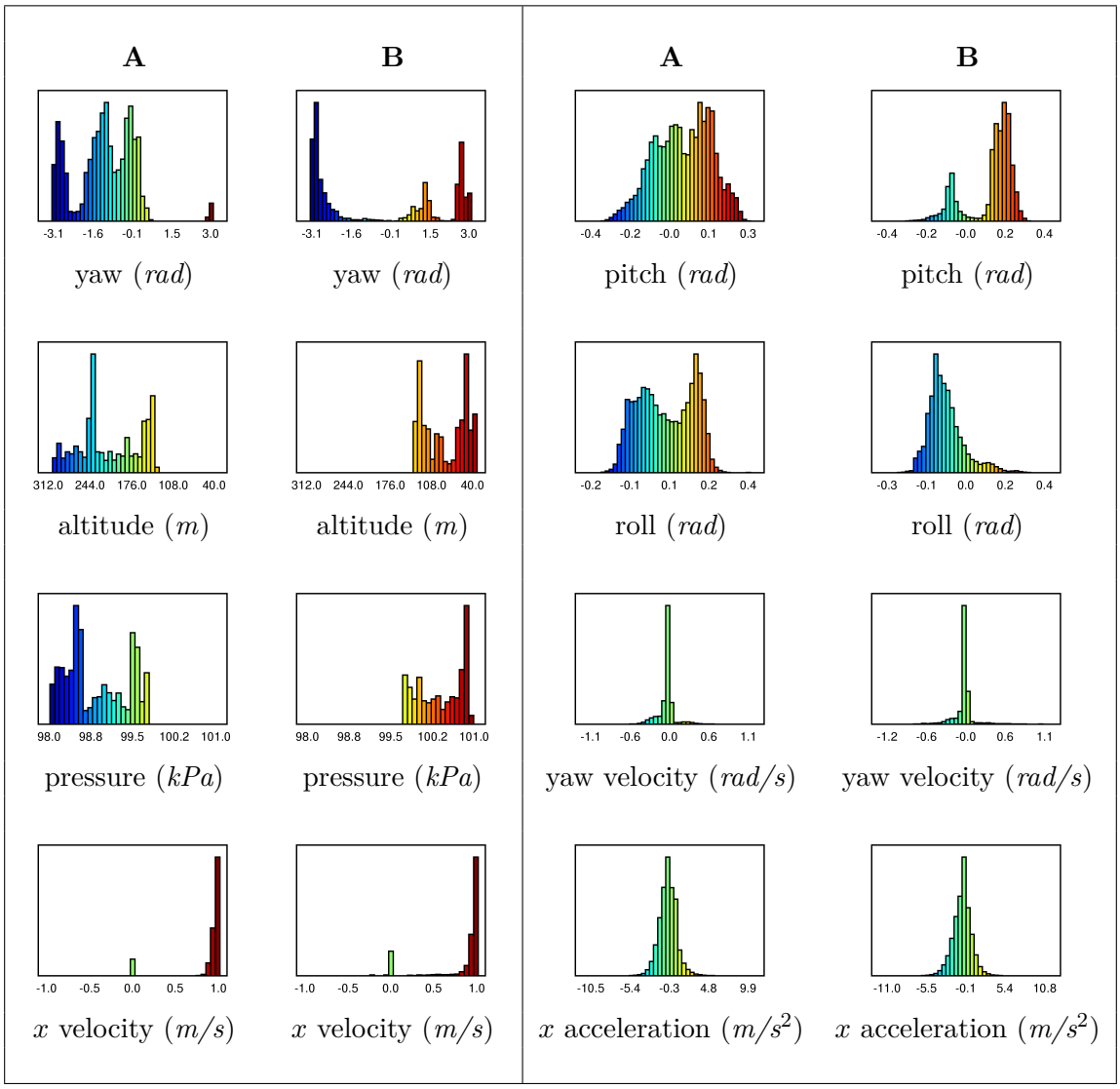


Figure 5.6: Histograms of sensor data from SFU Mountain Dataset

Chapter 6

Conclusion

6.1 Periodic Gestures

In this thesis, we proposed and demonstrated a vision system for long-range HRI: the first system to our knowledge that can locate and approach uninstrumented humans as small as 20 pixels tall, using only a low-resolution consumer camera. Once the robot has approached to within several meters, traditional close-range interaction techniques using other modalities become feasible [35].

We have published work on long-range HRI investigating video stabilization techniques to allow periodic gestures to be identified from moving cameras [2]. This enables detection during traversal for smoother behaviour, and permits the use of this method on aerial vehicles where remaining stationary is rarely an option. That work and the techniques introduced in this thesis improve on the maximum range of state of the art in HRI. Existing work requires humans to be on the order of 80 pixels tall for human detection methods such as existing HOG detectors [36], whereas our periodicity methods succeed for human heights of 20 pixels and below.

As mentioned in the discussion, methods for analyzing periodicity without compressing the color space to grayscale may reduce false positives caused by the tendency for grayscale to distort visible differences in color space. Other potential improvements include the use of machine learning as shown in [22] to distinguish robustly between human gestures and natural periodicity such as rustling foliage and flags blowing in the wind. Discriminating these from afar would help prevent the robot from leaving the area to investigate obvious distractors, thereby saving time and energy.

6.2 Trajectory Signals

We also proposed and demonstrated a behaviour sequence for a mobile robot to determine a human's intent to interact before approaching. This system is agnostic to sensor modal-

ity, and requires only an estimate of the relative trajectories of the two agents. We have also shown that the robot can obtain arbitrarily high levels of confidence by iterating the behaviour.

The system is trivially extensible to the multi-human case, because a single trajectory perturbation serves as a constant time behavioural probe for all humans under investigation at once. Extensions can be made for the multi-robot case, where the robots can coordinate their trajectories to most effectively disambiguate potential interested humans.

To our knowledge, this is the only work of its kind; no other approaches have been proposed for establishing joint attention by behaviour over large areas. Although we have no reference work to quantitatively compare with, we feel that such a behaviourally inexpensive probe that scales naturally to large numbers of agents and large interaction areas is unlikely to become obsolete in the near future.

6.3 Unification

We have described a system for detecting waving humans at long ranges, and an independent system for detecting intent to interact without completing an entire rendezvous. A waving detector can actually be used as the intercept detection mechanism, unifying the two main contributions of this thesis. This can be accomplished by exploiting a visual servoing trick known in psychology [37]: to intercept a moving object, simply keep it stationary in the field of view and as you move toward it, it will grow larger. Conversely for our case, if a detected gesture is stationary in the field of view and growing larger, then we can conclude we are on an intercept trajectory with the gesture source, and the trajectory behaviour can be applied.

6.4 Dataset

In addition to the HRI contributions, we presented a calibrated, synchronized, and ground-truth-aligned dataset of woodland trail navigation in semi-structured and changing outdoor environments. The data are highly challenging by virtue of the self-similarity of the natural terrain; the strong variations in lighting conditions, vegetation, weather, and traffic; and the three highly different trails. In the future we will expand this dataset by recording more traversals in different conditions. Notable desired conditions are autumn leaf colors, bare trees in winter, and the rare Burnaby winter snow. In contrast to most available datasets [32], the SFU Mountain Dataset contains the same environment under many different weather conditions, and the environment is largely unstructured, dynamic, and highly self-similar. To our knowledge, this dataset is the only available work that combines these features, and is thus an important contribution to help solve challenging navigation tasks in field situations.

Bibliography

[1] K. K. Kim, K.-C. Kwak, and S. Y. Ch, "Gesture analysis for human-robot interaction," in *Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference*, vol. 3, pp. 4 pp.–1827, Feb 2006.

[2] V. M. Monajjemi, J. Wawerla, R. T. Vaughan, and G. Mori, "HRI in the Sky: Creating and commanding teams of UAVs with a vision-mediated gestural interface," in *Proc. of Int. Conf. on Intelligent Robots and Systems*, 2013.

[3] D. Kim, J. Lee, H.-S. Yoon, J. Kim, and J. Sohn, "Vision-based arm gesture recognition for a long-range human-robot interaction," *The Journal of Supercomputing*, vol. 65, no. 1, pp. 336–352, 2013.

[4] S. Waldherr, R. Romero, and S. Thrun, "A gesture based interface for human-robot interaction," *Autonomous Robots*, vol. 9, no. 2, pp. 151–173, 2000.

[5] C.-C. Lien and C.-L. Huang, "Model-based articulated hand motion tracking for gesture recognition," *Image and Vision Computing*, vol. 16, no. 2, pp. 121 – 134, 1998.

[6] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 726–733 vol.2, Oct 2003.

[7] G. Rigoll, A. Kosmala, and S. Eickeler, "High performance real-time gesture recognition using hidden markov models," in *Gesture and Sign Language in Human-Computer Interaction* (I. Wachsmuth and M. FrÃűhlich, eds.), vol. 1371 of *Lecture Notes in Computer Science*, pp. 69–80, Springer Berlin Heidelberg, 1998.

[8] R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 416–416, IEEE Computer Society, 1998.

[9] X. Tong, L. Duan, C. Xu, Q. Tian, H. Lu, J. Wang, and J. Jin, "Periodicity detection of local motion," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 650–653, July 2005.

[10] P.-S. Tsai, M. Shah, K. Keiter, and T. Kasparis, "Cyclic motion detection for motion based recognition," *Pattern Recognition*, vol. 27, no. 12, pp. 1591–1603, 1994.

[11] M. Allmen and C. Dyer, "Cyclic motion detection using spatiotemporal surfaces and curves," in *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, vol. i, pp. 365–370 vol.1, Jun 1990.

[12] R. Polana and R. Nelson, "Detection and recognition of periodic, nonrigid motion," *International Journal of Computer Vision*, vol. 23, no. 3, pp. 261–282, 1997.

[13] R. Cutler and L. Davis, "Robust real-time periodic motion detection, analysis, and applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, pp. 781–796, Aug 2000.

[14] I. Laptev, S. Belongie, P. Perez, and J. Wills, "Periodic motion detection and segmentation via approximate sequence alignment," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1, pp. 816–823 Vol. 1, Oct 2005.

[15] F. Liu and R. Picard, "Finding periodicity in space and time," in *Computer Vision, 1998. Sixth International Conference on*, pp. 376–383, Jan 1998.

[16] Y. Ran, I. Weiss, Q. Zheng, and L. Davis, "Pedestrian detection via periodic motion analysis," *International Journal of Computer Vision*, vol. 71, no. 2, pp. 143–160, 2007.

[17] B. G. Quinn and E. J. Hannan, *The estimation and tracking of frequency*, vol. 9. Cambridge University Press, 2001.

[18] P. Borges, "Pedestrian detection based on blob motion statistics," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, pp. 224–235, Feb 2013.

[19] A. Briassouli and N. Ahuja, "Extraction and analysis of multiple periodic motions in video sequences," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, pp. 1244–1261, July 2007.

[20] S. Seitz and C. Dyer, "View-invariant analysis of cyclic motion," *International Journal of Computer Vision*, vol. 25, no. 3, pp. 231–251, 1997.

[21] J. Sattar and G. Dudek, "Where is your dive buddy: tracking humans underwater using spatio-temporal features," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pp. 3654–3659, Oct 2007.

[22] M. Takahashi, K. Irie, K. Terabayashi, and K. Umeda, "Gesture recognition based on the detection of periodic motion," in *Optomechatronic Technologies (ISOT), 2010 International Symposium on*, pp. 1–6, Oct 2010.

[23] M. Monajjemi, J. Bruce, A. Sadat, J. Wawerla, and R. Vaughan, "UAV, do you see me? Establishing mutual attention between an uninstrumented human and an outdoor UAV in flight," in *IEEE Int. Conf. on Intelligent Robots and Systems*, 2015.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *KDD*, vol. 96, pp. 226–231, 1996.

[26] M. J. Matarić, J. Eriksson, D. J. Feil-Seifer, and C. J. Winstein, "Socially assistive robotics for post-stroke rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 4, p. 5, 2007.

[27] T. Fong, C. Thorpe, and C. Baur, "Collaboration, dialogue, human-robot interaction," in *Robotics Research*, pp. 255–266, Springer, 2003.

[28] J. Bruce, J. Wawerla, and R. Vaughan, "Human-robot rendezvous by co-operative trajectory signals," in *10th ACM/IEEE International Conference on Human-Robot Interaction, Workshop on Human-Robot Teaming*, 2015.

[29] S. Pourmehr, V. M. Monajjemi, R. T. Vaughan, and G. Mori, "You two! Take off! Creating, modifying and commanding groups of robots using face engagement and indirect speech in voice commands," in *Proc. of Int. Conf. on Intelligent Robots and Systems*, 2013.

[30] P. Fiorini and Z. Shiller, "Motion planning in dynamic environments using velocity obstacles," *The International Journal of Robotics Research*, vol. 17, no. 7, pp. 760–772, 1998.

[31] J. Bruce, J. Wawerla, and R. T. Vaughan, "The SFU Mountain Dataset: Semi-structured woodland trails under changing environmental conditions," in *Workshop on Visual Place Recognition in Changing Environments at the IEEE International Conference on Robotics and Automation (ICRA'15 workshop), Seattle, WA, USA*, (Seattle, WA, USA), May 2015.

[32] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. of Robotics Research*, 2013.

[33] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. of Robotics Research*, vol. 27, no. 6, 2008.

[34] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Int. Conf. on Robotics and Automation (ICRA)*, IEEE, 2012.

[35] S. Pourmehr, J. Bruce, J. Wawerla, and R. Vaughan, "A sensor fusion framework for finding an HRI partner in a crowd," in *IEEE Int. Conf. on Intelligent Robots and Systems, Workshop on Designing and Evaluating Social Robots for Public Settings*, 2015.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[37] M. Lenoir, E. Musch, M. Janssens, E. Thiery, and J. Uyttenhove, "Intercepting moving objects during self-motion," *Journal of Motor Behavior*, vol. 31, no. 1, pp. 55–67, 1999.