Tiny People Finder: Long-Range Outdoor HRI by Periodicity Detection

Jake Bruce, Valiallah (Mani) Monajjemi, Jens Wawerla and Richard Vaughan
Autonomy Lab, Simon Fraser University
{jakeb, mmonajje, jwawerla, vaughan}@sfu.ca

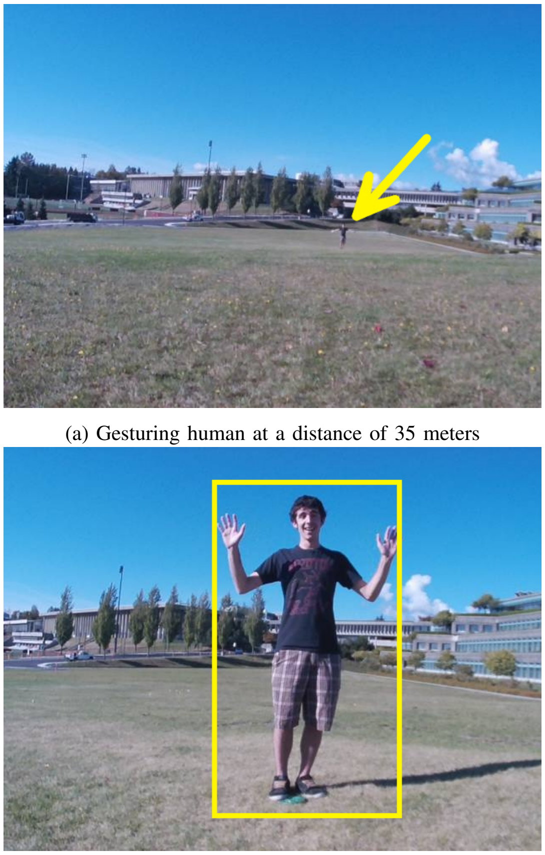
***Abstract*— We present a novel method for detecting waving humans at long ranges in outdoor environments, using a consumer video camera on a mobile ground-based robot. The proposed algorithm analyzes the average pixel intensity of motion-containing regions in an image stream, identifying those regions which show a strong periodic signal in the frequency range of human waving gestures. The system achieves robustness to periodic false positives such as waving trees and flags by using behavior to confirm or reject the detection at close range. In real-world experiments we show that a robot equipped with a low-resolution consumer camera is able to approach a single waving human from a starting position up to 35 meters away, even in the presence of non-human periodic distractors such as foliage.**

I. INTRODUCTION

Consider the following situation: you're standing on a hill looking down into a crowd of people around a hundred meters away, attempting to find a friend. You know this friend well; you'd recognize her face immediately if she approached you, but from this distance it is hard to make out the details of the crowd, 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 sophisticated 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.

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, and against moving and cluttered backgrounds (see Fig. 1). 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 and confirms or rejects the candidate at short range. 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.



(a) Gesturing human at a distance of 35 meters

(b) Human approached and detected

Fig. 1: (a) long-range view of waving human using a 640×480-pixel camera, and (b) short-range view after approach, with human-detection bounding box.

The contributions of this paper are (1) a long-range periodic gesture detection algorithm that reliably identifies periodic motions at distances of up to 35 meters using a low-resolution camera, and (2) a robot behavior that provides robustness to false positives of long-range object detectors by confirming or rejecting the candidate detection at close range. The HRI system in this paper is the first to our knowledge that can locate and approach uninstrumented gesturing humans at ranges of up to 35 meters in indoor and outdoor environments, using only monocular camera sensing.

The rest of the paper is organized as follows: section II reviews related work. We describe how periodic regions are identified from live camera footage in section III. The behavior of the robot is described in section IV followed by experiments and results in section V. In section VI, we conclude the paper and reflect on future extensions of this work.

II. 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, as does the performance of optical flow-based techniques [8], [9].

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 artificial periodicity due to small errors in feature tracking and image alignment methods at long ranges, 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 applied to gesture recognition 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.

III. Periodic Motion

The proposed algorithm for detection of periodic image regions can be described in four stages: *A)* motion identification to limit computation to moving image regions, *B)* constructing a time series of average pixel intensity on a per-region basis, *C)* identification of periodic signals in the desired frequency range for human waving, and *D)* clustering periodic regions into large-scale bounding boxes for output to the behavior system. This section describes each stage in detail.

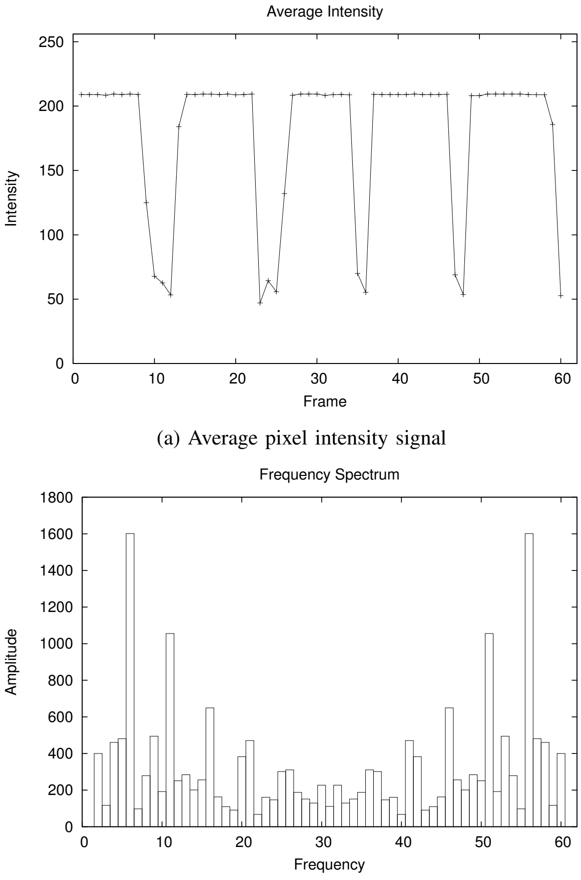
A. Identifying Motion

In order to detect small motions on the order of 20 pixels in height we require that the robot be stationary, since even small errors in image registration methods can cause false positives and wash out periodic signals. For the purpose of reducing computation time, we limit processing to regions that contain visual motion by constructing a foreground mask out of the pixel-wise difference image between frames (see Fig. 3a). The pixels of this mask are aggregated over 30 frames to produce a motion silhouette mask. Pixels that belong to the foreground mask are used in constructing the intensity time series on the basis of which to discriminate periodic motion. Assuming the robot is stationary simplifies this component: in future work we will remove this constraint.

B. Intensity Time Series

Once foreground pixels are identified, overlapping square bounding boxes are marked as potential periodic regions. A box belongs to this collection only if at least one of the pixels it contains belongs to the foreground mask that indicates a nonzero difference between consecutive frames. Since we are interested in detecting people as small as 20 pixels in height, we use bounding boxes 10 pixels on a side, and we overlap these boxes by half along each axis.

A weighted average grayscale intensity of the pixels in each box is computed using a Gaussian kernel centered on the middle of the bounding box. This non-uniform weighting reduces edge effects of the bounding boxes, and ensures that a pixel that only moves inside one bounding box still produces variation in its box's average intensity. The system is not sensitive to the parameters of the weight surface, but a steeper Gaussian results in pixels nearer the center of each



(a) Average pixel intensity signal

(b) Frequency spectrum with DC component removed

Fig. 2: Intensity signal and frequency spectrum extracted from a periodic waving gesture. The periodicity P_F of this signal is 0.61.

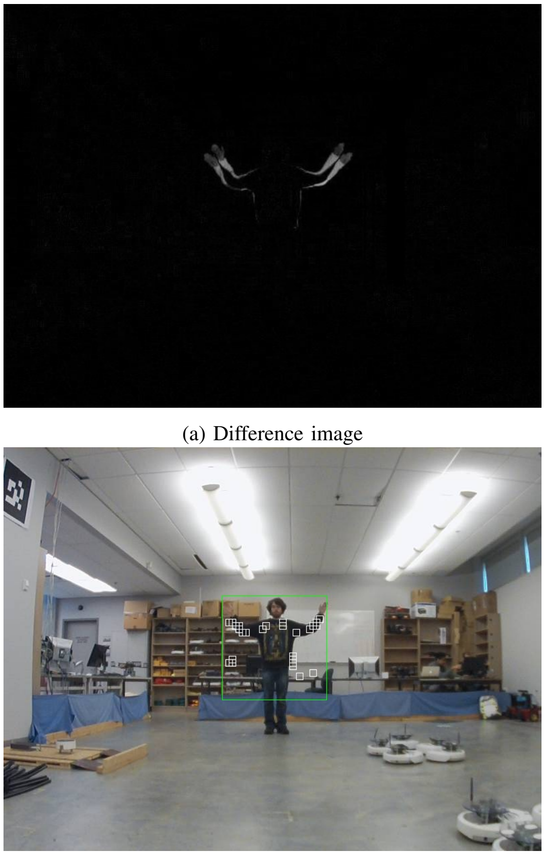
box having more impact on the average. We use a Gaussian kernel with the following covariance:

$$\mathbf{\Sigma} = \begin{pmatrix} 25 & 0 \\ 0 & 25 \end{pmatrix}$$

These weighted averages are stored in a 2-second buffer for each bounding box: a typical series is shown in Fig. 2a. 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.

C. 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 must discriminate 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. 2b). 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:



(a) Difference image

(b) Periodic regions in white and cluster bounding box in green

Fig. 3: Stages of periodic motion identification for a waving human at close range. (a) shows the frame-to-frame difference image used to limit computation to motion-containing regions, and (b) shows the periodic regions and bounding box formed by the algorithm in this paper.

$$P_F = \frac{\sum_i F_{iw} - \sum_i F_{iw+w/2}}{\sum_i F_{iw} + \sum_i F_{iw+w/2}} \tag{1}$$

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.

We consider a signal to represent a potential gesture if $P_F > 0.25$ and w is between 1Hz and 2Hz, as humans tend to wave at approximately this rate. In addition, we only consider signals with fundamental amplitude $F_w > 10$, 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.

D. Clustering

Given a collection of 10×10-pixel regions flagged as positive for periodic motion, the system forms large-scale bounding boxes (see Fig. 3b) to identify multiple sources of motion if present. We use the SciPy implementation [23] of the DBSCAN algorithm [24] 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 are not passed on to influence behavior. We use $\epsilon = 15$ pixels and $\delta = 3$, chosen to form sensible clusters at both short and long ranges.

IV. Robot Behavior

A detector that only functions while the robot is stationary presents a challenge for defining behavior. We cannot simply servo to the detection, as camera movement can produce apparent periodicity in non-periodic scenes. The relative 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.

We estimate the distance to a gesture by assuming a pinhole camera with vertical resolution h_p and vertical angular field of view h_θ, and a typical waving gesture size g_m of 1 meter. Given the size g_p of a detected gesture in pixels, the distance estimate d is computed as:

$$d = \frac{g_m}{\sin(\frac{g_p}{h_p} h_\theta)} \tag{2}$$

This estimate agrees with measured values to within 10 meters, at the distances tested in our experiments. This is appropriate for our system, since other techniques such as face or torso detection may be used at 10 meters for a closer approach and subsequent interaction. False positive periodic motions tend to occur high up in tree foliage, on flag poles and in other unreachable locations, so the approach stage functions as a rough filter to reject many long-range detections as they move out of view.

We define an approach as successful once the estimated distance to the detection is within 10 meters and the detection is confirmed with the well-established histogram of oriented gradients (HOG) algorithm [25]. We use the OpenCV HOG implementation [26], trained for human detection. If a gesture is detected with an estimated distance of less than 10 meters and is not confirmed by the human detector, or if no gesture is detected, the robot rotates through the angular field of view of its camera and begins another stationary scan. If the estimated distance to the gesture is greater than 10 meters, the robot drives half the estimated distance toward the detection, up to a maximum of 15 meters, and begins another scan.

In the case of multiple detections, if the scan phase was preceded by an approach phase, the robot approaches the detection closest to the center of its field of view. If the scan was preceded by a rotation, the robot approaches the detection with the largest bounding box and thus the least estimated distance. This ensures that the robot will stick to its previous target when possible, and investigate the closest candidates first.

V. Experiments

We evaluate the proposed system using a Husky A200 ground-based robot built by Clearpath Robotics. Excluding emergency-stop behavior, the only active sensor for these experiments is a consumer 640×480 resolution monocular Axis camera mounted on the front of the robot (see Fig. 5), providing color video at 30 frames per second. The robot includes an onboard computer with 8GB of memory and a four-core Intel Core i5 2.40GHz 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. 4 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 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. 1 until the robot stops within 10 meters, which is considered a

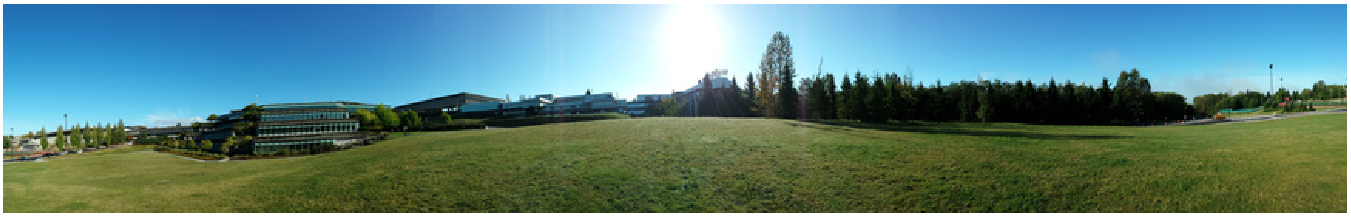


Fig. 4: Outdoor experimental setting on SFU campus.

Trial		Orientation (°)	Time (s)	Result
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 I: Experimental results with robot in different initial orientations. *S* indicates a successful approach, and *F* indicates an approach that did not arrive within 10 meters of the human after 180 seconds.

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 I, and approach trajectories are shown in Fig. 6.

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. 6a, 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.

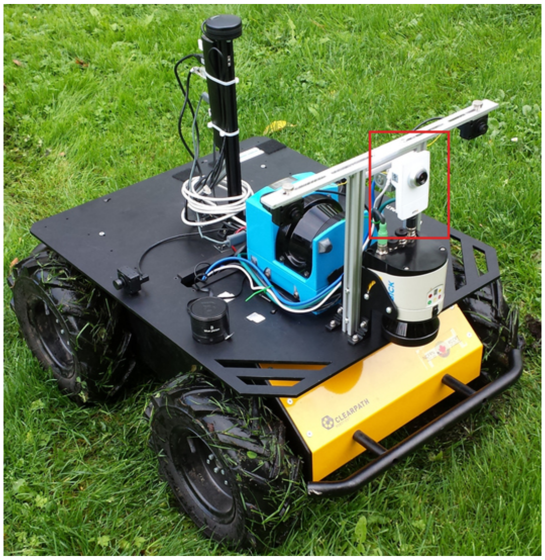
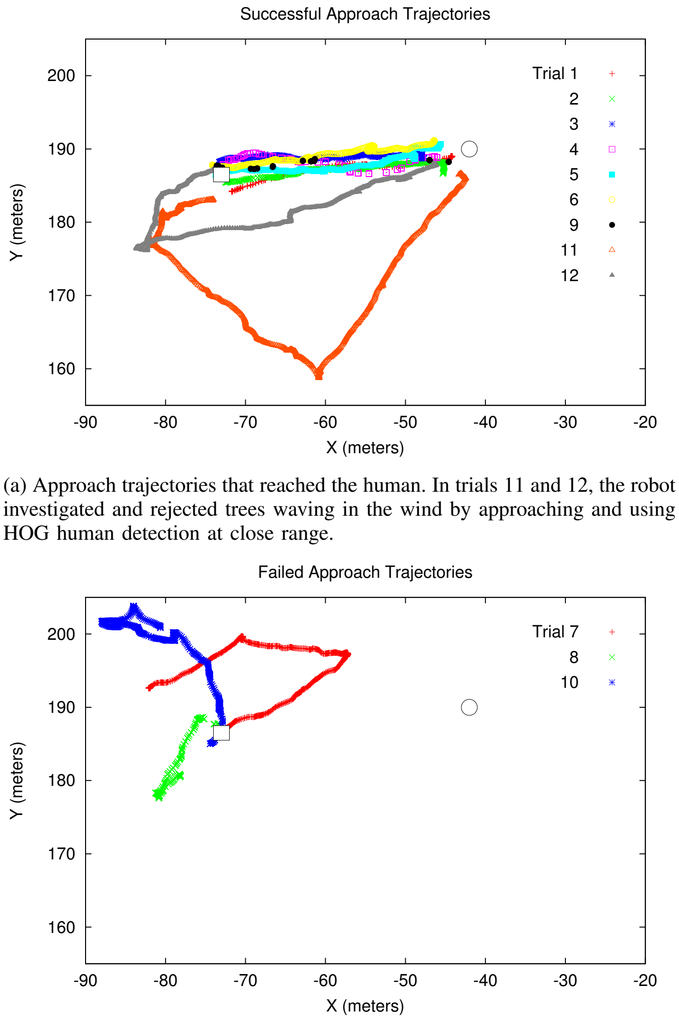


Fig. 5: Husky A200 robot used in these experiments, with monocular Axis camera outlined.



(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.

Fig. 6: Experimental approach trajectories. Square indicates robot start position and circle indicates human location.

VI. Conclusions and Future Work

We propose and demonstrate a vision system for long-range HRI: the first system to our knowledge that can locate and approach uninstrumented humans from ranges up to 35 meters, using only a low-resolution consumer camera. Once the robot has approached to within 10 meters, traditional close-range interaction techniques become available using monocular and stereo cameras, laser rangefinders, and voice-based interaction.

Future work on long-range HRI will include the investigation of video stabilization techniques to allow periodic gestures to be identified from moving cameras. This will allow detection during traversal for smoother behavior, and will permit the use of this method on aerial vehicles where remaining stationary is rarely an option.

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. The system in this paper tends to reject detections of this sort since they usually do not occur at approachable locations at ground level and quickly leave the field of view of the camera, but discriminating these from afar would help prevent the robot from leaving the area to investigate obvious distractors.

Although we did not use it for the purpose of these experiments, our robot vehicle includes a stereo camera assembly. Several improvements could be made to the long-range interaction system with the use of stereo disparity information, including estimating the distance to the target during the approach, and rejecting detections that lie far from the ground plane.

ACKNOWLEDGMENT

Thanks to Jacob Perron for assisting with experiments. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) through the NSERC Canadian Field Robotics Network (NCFRN).

REFERENCES

[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. Monajjemi, J. Wawerla, R. Vaughan, and G. Mori, "HRI in the sky: Creating and commanding teams of UAVs with a vision-mediated gestural interface," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pp. 617–623, Nov 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. Frhlich, 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] E. Jones, T. Oliphant, P. Peterson, *et al.*, "SciPy: Open source scientific tools for Python," 2001–. [Online; accessed 2014-09-19].

[24] 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.

[25] 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.

[26] G. Bradski, "OpenCV: the open source computer vision library," *Dr. Dobb's Journal of Software Tools*, 2000.