UAV, Do You See Me? Establishing Mutual Attention Between an Uninstrumented Human and an Outdoor UAV In Flight

Mani Monajjemi, Jake Bruce, Seyed Abbas Sadat, Jens Wawerla and Richard Vaughan*

***Abstract*— We present the first demonstration of establishing mutual attention between an outdoor UAV in autonomous normal flight and an uninstrumented human user. We use the familiar periodic waving gesture as a signal to attract the UAV's attention. The UAV can discriminate this gesture from human walking and running that appears similarly periodic. Once a signaling person is observed and tracked, the UAV acknowledges that the user has its attention by hovering and performing a "wobble" behavior. Both parties are now ready for further interaction. The system works on-board the UAV using a single camera for input and is demonstrated working reliably in real-robot trials.**

I. Introduction

Flying robots have many obvious applications and are developing very rapidly. Our work on Human-Robot Interaction (HRI) aims to create interfaces to robots that resemble those we use with humans and other animals, using gestures, body pose and speech. We avoid using computer-based interfaces because they require instrumenting the user, preventing the robot from interacting ad-hoc and spontaneously with an arbitrary person. Building on our previous work on direct HRI with small, hovering indoor Unmanned Aerial Vehicles (UAVs) [1] we present the first demonstration of establishing mutual attention between an uninstrumented human user and an autonomous, self-contained outdoor UAV in normal, non-hovering flight.

We use the familiar periodic double-arm-waving gesture pictured in Figure 1 as a signal to attract the UAV's attention. The UAV can discriminate this gesture from human walking and running that appears similarly periodic. Once a signaling person is observed and tracked, the UAV acknowledges that the user has its attention by hovering and performing a distinctive "wobble" behavior. Both parties now know that they have the attention of the other, and are now ready for further interaction. The system works on-board the UAV using a single camera for input and is demonstrated working reliably in real-robot outdoor trials.

The problem of detecting and tracking humans from a moving camera platform is non-trivial, and is a current research problem. It is essential for pedestrian detection in self-driving cars, and for automated surveillance from drones. We [1] and others [2]–[4] have previously demonstrated people-tracking for close-up HRI with small UAVs. Expanding this work outdoors and to a UAV that is flying rather than hovering, we find that the person is represented by only a very few pixels in the image and may be in-view only briefly.

* Autonomy Lab, School of Computing Science, Simon Fraser University {mmonajje, jakeb, asadat, jwawerla, vaughan}@sfu.ca

Fig. 1: Our system in action, showing scale. A human is waving to a UAV.

Variable lighting conditions and fast camera motion also contribute to the challenge. Our approach is to use simple, fast computer vision methods that require no training and can run in real-time on-board the UAV.

The periodic waving gesture is designed to be highly salient to an observer: it makes the user appear larger, and fast periodic motions are relatively rare in outdoor scenes. Thus it is amenable to computer vision detection.

The arm-waving signal is also very familiar, and we informally suggest that this is a natural way to attract the attention of any human, animal or robot that is looking for you. In addition to being familiar and easy to perform, the user's intention to attract the robot can be correctly interpreted by human observers. Again, informally, we suggest that this behavior does not need to be taught to users, so our system could work in search and rescue scenarios where the subject has no robot training. In wilderness survival guides, this behavior is suggested as an effective signaling method to attract attention (e.g. [5]).

The robot signals back to the user with a high-frequency periodic "wobble" behavior as an approximation of the "wing-waggle" behavior conventionally used by fixed-wing aircraft pilots to show they have observed a person on the ground. Informally, this is readily perceived by the user as a confirmation of being attended to, probably because it looks deliberate yet distinct from the normal control motions of the vehicle.

Our vision system provides detection and tracking of candidate *humans-that-want-to-interact* for moving UAVs. It occasionally gives false positives and negatives, so should be used as part of a closed-loop system whereby the UAV has suitable failure/retry behaviors to achieve real-world robustness [6].

The contributions of this paper are (i) a description of a

simple but fast approach to detect waving gestures based on a combination of well-known computer vision methods; (ii) a demonstration of the first fully autonomous UAV detecting the intention to interact from an uninstrumented person with all computation performed on-board. (iii) a complete implementation available online.

II. Background

Interfaces to control UAVs can broadly be classified into two groups. Those that use conventional instrument-based Human-Computer Interfaces [7], [8] and direct and uninstrumented interfaces mostly based on computer vision techniques. Of the uninstrumented interfaces, a few have demonstrated fully integrated human-UAV interaction systems. Lichtenstern et al. [2] described a system in which gestures observed by a UAV carrying a Microsoft Kinect (indoor, active RGB-D) sensor are used to control other UAVs. Naseer et al. [3] developed an autonomous system that enables a single quadrotor to follow a human and respond to hand gestures using an active RGB-D sensor with vision-based ego-motion cancellation. Costante et al. [4] developed a person-specific gesture-based interface to command a UAV using monocular vision.

All these systems use vision-based body or face detectors to find the region of interest (ROI) in the robot's field of view for tracking the human and/or performing gesture recognition. The RGB-D based solutions are not applicable to outdoor settings or long distances because sunlight overwhelms the projected infrared structured light field. Furthermore, state of the art human-detectors are too computationaly intensive to run in real-time on a CPU and are unreliable when the person is distant (< 30 pixels tall) [9]. As a result, all existing approaches have been applied to close range interaction scenarios in indoor environments only. Our human-UAV interaction system is the first to work outdoors, while the UAV is translating rather than hovering, and over relatively long distances ($> 10m$) when the human occupies less than 5% of the image.

Instead of directly using human-detectors to find and track the human in the UAV imagery, it is possible to find and track objects using motion or other salient object detection techniques (also known as foreground object segmentation). Sokalski et. al [10] developed a system that combines contrast features, mean shift segmentation and multichannel edge features to detect static salient regions in UAV imagery. Rodriguez et. al [11] developed a real-time system to detect and track multiple moving objects from a UAV by constructing an artificial sparse optical flow field from estimated camera motion. Discrepancies between the real and artificial flow fields characterize moving objects. Camera motion is estimated by a monocular visual SLAM algorithm. Siam and Elhelw [12] developed a similar system to track multiple moving objects from a UAV by clustering feature points which are outliers in camera motion estimation step. The camera motion estimation is done by finding the homography transform between consecutive frames using tracked feature points. Kimura et al. [13] applied multi-view geometry constraints (epipolar constraint and flow-vector bound) to tracked feature points between consecutive frames in order to detect moving objects from an airborne UAV. Our approach is similar to [11]–[13] in the sense that we also rely on explicit camera motion estimation and motion saliency to detect regions of interest. However since the arm waving gesture is not a strong motion cue from a distance, we also integrate tracked feature points to find salient objects.

To detect a dual arm waving gesture given a sequence of tracked ROIs, it is possible to apply two approaches: human activity recognition techniques and periodicity analysis. Human activity recognition is an active and vast area of research. Although there exist many promising human activity recognition algorithms, most are far from real-time on current hardware [14]. For the specific task of action recognition from distance, the computation time is dominated by the need for precise tracking (e.g [15], [16]) or an expensive motion feature extraction step (e.g [16], [17]).

Detecting and analyzing periodicity in image sequences has been explored in the human activity recognition community to classify cyclic human actions (e.g. walking or waving). Some earlier work [18], [19] relies on detection and tracking of specific points on the human body to detect periodicity. This method is not practical in our setting, since tracking feature points reliably on a small moving object from distance is not feasible. Methods based on temporal changes in individual pixel intensities use frequency domain analysis (e.g [20]), periodicity metrics such as periodograms (e.g [21]) or self-similarity (e.g [22]) to detect periodicity in regions of interest. Since these approaches require precise frame alignment, they become computationally expensive and inaccurate in the presence of tracking errors. Another approach to detect periodicity is to consider the ROI as a whole and perform frequency domain analysis of either its trajectory [23], [24] or mean pixel motion [25]. We found these methods less sensitive to tracking errors and thus a better fit for detecting periodicity using a moving camera. Similar to [25] we perform frequency domain analysis on the average motion per pixel of each ROI. Periodic motion detection has previously been successfully applied to outdoor human robot interaction: in Sattar et. al [26], an underwater robot follows a human diver by tracking the periodic motion of the diver's brightly-colored fins. The robot also uses blob tracking to compensate for tracking errors. This work is different, since we perform periodicity detection on-board a UAV and does not rely on strong color or other appearance priors.

III. Method

To detect hand waving signals from a flying platform, we first estimate the camera motion between consecutive frames—and hence the robot's ego motion—by tracking feature points between frames. It is also used later in the pipeline to estimate each salient object's movement over time with respect to a fixed reference frame. To find salient objects, we first cluster tracked feature points using a fast non-parametric clustering algorithm. Clusters are first pruned based on size

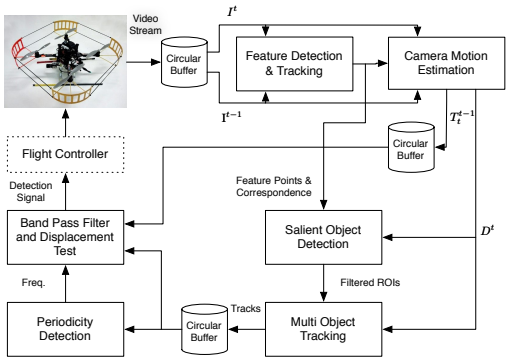

Fig. 2: Block diagram of the system. Refer to Sec. III for the definition of each variable. Except flight controller, other components run on the vision processing unit (Sec. III-E).

and motion characteristics and then tracked using a bank of Kalman filters. For each tracked object's ROI, we calculate average motion per pixel and use discrete Fourier transform and a statistical test to estimate the dominant frequency component of that signal. Tracks that are (i) stationary in the global reference frame and (ii) show periodicity in a specific frequency band are classified as positive detections. Figure 2 shows the overview of the detection pipeline. In the following sections we describe each step in more detail.

Throughout this section we introduce parameters for each component, but we defer the discussion on how to set these parameters to Sec. III-D. Since our approach is a temporal method, we extensively use circular buffers of fixed size N to store a rolling history of different entities. We use the terms "sampling period" and "sequence length" interchangeably to denote N.

A. Camera Motion Estimation and Ego Motion Cancellation

Our camera motion estimation and stabilization method is based on techniques developed for visual odometry with monocular cameras [27], [28]. To estimate the camera motion at time t, the input frame is first converted to a grayscale image I^t. Then, we detect FAST corners [29] in I^t and store them in a list $F^t = \{f_i^t\}$. To limit computation time in scenes with large number of strong corners, we limit the number of stored feature points to N_F^{max}. We use a pyramidal implementation of the Lucas Kanade optical flow algorithm [30] to find matches between F^{t-1} and F^t. If the number of matches exceed threshold N_F^{min}, we fit a full perspective motion model on the optical flow field using least median of squares regression. Otherwise, we consider motion estimation as failed for the current frame and does not execute other components of the pipeline. After pruning outliers, the result is further refined using Levenberg-Marquardt non-linear optimization. The resulting transform T_t^{t-1} from frame t to frame $t-1$ is stored in a circular buffer. To cancel out the ego motion of the vehicle, we warp I^t into I^{*t} using the inter-frame transform T_t^{t-1}. We calculate the absolute image difference of I^{*t} and I^{t-1}, adaptively threshold the result and pass it through a low-pass filter for smoothing and suppressing transient errors. We call the resulting image D^t "camera independent inter-frame motion image".

B. Salient Object Detection and Tracking

In order to detect salient objects in each frame, we combine two cues: camera independent inter-frame motion and spatial density of feature points in that frame. Motion fields (similar to D^t) have been widely used by researchers to segment and track moving objects from mobile cameras (e.g [22]). However the segmentation quality is heavily dependent on the quality of the camera motion estimation, type of background and size of targets. For our specific application, the size of the target (two moving arms) can be as small as 10×10 pixels, the background is usually complex in outdoor settings, and there is often inevitable camera stabilization error. Camera stabilization error causes false motion blobs in areas with non-homogeneous background as well as around objects with feature points not lying on the ground plane. For those reasons, we found that the motion field alone is insufficient to segment hand waving motion from a distance.

Our approach for detecting salient moving objects is based on detecting dense clusters of feature points in motion salient areas of the image. We first use DBSCAN [31], a fast non-parametric density based clustering algorithm to detect dense clusters of feature points in the frame. As we will show in Sec. IV, it runs in real-time when clustering hundreds of feature points. DBSCAN only relies on two parameters: the maximum inner cluster distance ϵ and the minimum number of feature points per cluster N_c^{DBS}. For each cluster, elements that have zero motion are discarded ($D^t(X_{f_i^t}, Y_{f_i^t}) = 0$). Next a minimum axis-aligned bounding box is fitted to the remaining members. A post-pruning step filters out clusters that are smaller than $S_{min} = W_{min} \times H_{min}$, larger than $S_{max} = W_{max} \times H_{max}$ or have small average motion per pixel value. Given a bounding box B^t, the average motion per pixel ($D_{avg}^{B^t}$) is calculated as follows:

$$D_{avg}^{B^t} = \frac{\sum_{(x,y)\in B} D^t(x,y)}{W_B \times H_B} \tag{1}$$

We use a bank of Kalman filters with a constant-acceleration motion model to track the state of each cluster (position, velocity and size) over time. To cancel out the effect of ego motion in state transition, the state is warped using T_t^{t-1} before each Kalman prediction step. In other words, at time t, the previous state of each track is first transformed to the current frame's coordinate system using the inverse of the estimated camera motion, then the Kalman prediction step is applied. To associate observations to tracks we use the Hungarian matching algorithm with extensions proposed in [32]. Tracks without any associated observation are deleted after a timeout period.

To differentiate between stationary periodic actions such as hand waving gestures, and non-stationary periodic ones such as walking we calculate the camera independent displacement of each track over the sampling period δ_{t-N-1}^t. To determine this value we first need to calculate the camera motion over the whole period:

$$T_t^{t-N-1} = \prod_{i=t-N}^{i=t} T_i^{i-1} \tag{2}$$

Then we remove the effect of camera motion from the position of each tracked object ($P^t = [x^t y^t]^T$):

$$P_s^t = T_t^{t-N-1} P^t \tag{3}$$

The Euclidean distance between P_s^t and P^{t-N-1} in image space is the camera independent displacement of the tracked object over the sampling period.

C. Periodicity Detection

To detect periodicity we perform frequency domain analysis on each track's average motion per pixel (Eq. 1) over the sampling period. We chose this measure since it is fast to calculate and unlike pixel intensity based measures, does not require perfectly aligned tracks. The latter is important, because we found precise tracking to be difficult to achieve in real-time under fast camera motion in flight and when the tracked object is non-rigid.

For each track, the average motion per pixel signal $D_{avg}^t(t)$ is first de-trended and windowed with the Hann function. Using a discrete Fourier transform, we calculate the power spectrum of the signal and find its maximum normalized power component. If A_k where $k \in K = \{1..\frac{N}{2} - 1\}$ denotes the positive half of the energy spectrum, the maximum normalized component is calculated as follows:

$$A_M = \frac{arg\,max_{k \in K} A_k}{\sum_{k \in K} A_k} \tag{4}$$

To test if A_M is statistically significant and thus is the dominant frequency of the signal, we apply approximation to Fisher's exact test proposed by [33]. If A_M passes this test with confidence greater than 99.5%, we consider the track as periodic with frequency $f = \frac{k \times fps}{N}$.

If a track's dominant frequency is between f_{min} and f_{max} with small camera independent displacement $\delta_t^{t-N-1} < \delta_{max}$, we classify that track as stationary, periodic gesture.

Figure 3 shows the effect of each component in the pipeline to detect stationary periodic objects on a sample sequence from the ARG dataset (Sec. IV-A).

D. Tuning Parameters

Our system is sensitive to two of the parameters described so far: the maximum inner cluster distance of DBSCAN (ϵ) which controls the size of objects of interest in the scene and the video frame-rate (FPS) that limits the accuracy of the periodicity detection component. Setting FPS is trivial because it is known in advance. We manually tuned ϵ for specific experiments. However, it is possible to tune this parameter automatically given the height above ground at which the UAV is flying, camera intrinsics and a prior on the size of objects of interest (people in our case). We set N_F^{max} and N_F^{min} to 500 and 10 feature points respectively. For smaller input sizes, we reduce this number. The sequence length N is set to four times the FPS value (100 to 120) to capture a few periods of the gesture. The parameters of band pass filter for periodicity detection is set to $f_{min} = 0.9$ Hz and $f_{max} = 3.0$ Hz to include the frequency range of human waving gestures. To reject small or large bounding boxes we set $S_{min} = 5 \times 10$ and $S_{max} = 100 \times 200$ pixels. Similar to ϵ, these two parameters can be inferred automatically. Finally we set the threshold to segment stationary and non-stationary tracks (δ_{max}) to 30 pixels.

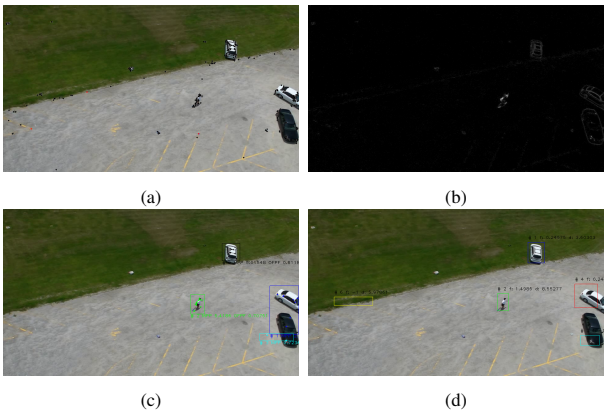

Fig. 3: The output of each component of the pipeline running on a sample from ARG Aerial dataset (a) Tracked feature points and correspondence (b) Camera independent inter-frame motion image (c) Salient objects (d) Tracks

E. Platform and Implementation

We run the entire system on board an Asctec Pelican quadrotor to create a fully autonomous system capable of establishing joint attention with an unistrumented human in outdoor settings. The pipeline runs on a small form factor PC with a dual core 4th generation Intel Core i7 CPU and 8GB of RAM. To capture images, we use a Point Grey Firefly MV color camera mounted on an actively stabilized gimbal. The Firefly MV is a global shutter camera which captures 640×480 color images up to 60 frames per second. The on-board computer controls the UAV by sending position and velocity commands to the flight controller. The total weight of the entire vision processing system (camera, small form factor PC and battery) is 400 grams. The pipeline is implemented in C++ and relies on an optimized build of OpenCV [34]. [1]

IV. Experiments

In this section we first report the performance of the proposed method on two human action datasets, then we describe our experimental setup to demonstrate the effectiveness of this method in a human flying robot interaction scenario. As discussed in Sec. III-D, two parameters have to be tuned for a specific data source: Framerate (FPS) and the maximum inner cluster distance of DBSCAN algorithm (ϵ). We tuned the latter for each dataset to achieve good performance.

[1]All source code and configurations used to generate the results in this paper are available for download at `http://autonomylab.org/obzerver/`. The commit hash of the code used to generate the results begins `1fc6bd8`.

TABLE I: Properties of video streams, parameters used for each experiment and average execution time per frame (Sec. IV-C)

Setup	Picture Size	FPS (Input)	ϵ	Avg. Exec. Time per frame (ms)
KTH	160×120	25	0.2	3.136
ARG	960×540	30	0.03	27.84
UAV	640×480	30	0.2	31.65

A. Datasets

We tested the system on two human action datasets to evaluate the precision and performance of the proposed method in detecting hand waving gestures and rejecting periodic distractions such as walking and running people.

The first dataset is the *KTH human action dataset* [35], which contains six actions performed by 25 actors. The camera is static and the background is homogeneous. Each action is performed four times by each actor in four different scenarios: static homogeneous background (SHB), SHB with scale variation, SHB with different clothes and SHB with lighting variations. The second dataset is the *UCF-ARG (University of Central Florida-Aerial camera, Rooftop camera and Ground camera)* [36]. We use the aerial component of the dataset which was recorded from a remote controlled helium balloon. It consists of 10 actions performed four times (in different directions) by 12 actors in an open parking lot. This dataset contains a set of challenges: fast and sudden camera motion in almost all video frames, shadows, variation in scale and clothing and small size of people in this dataset which occupy less than 5% of the whole 960×540 image. Table I lists the properties of the video stream in each dataset as well as the parameters we used to evaluate the system.

First we report the performance of our approach on detecting human hand waving gestures on these datasets. Table II summarizes the true detection rate, false positive rate and miss rate of the vision pipeline when applied to hand waving gesture subset of these datasets. A detection is considered correct if the detected bounding box is stationary, overlaps with the upper body of the actor and includes at least one hand. If the system detects a bounding box which is stationary but does not overlap with the body it is considered as a false positive. Non-stationary detections as well as no detections are considered as misses. The detection rate on KTH and ARG datasets are 78% and 56.25% respectively. Although, the false positive rate is low for both datasets (0% and 4.15%), the miss-rate is the major deficiency. We observe that salient object detection and tracking errors due to scale changes (KTH), small objects and fast camera motion (ARG) are the main causes of the high miss rate. Since the input video length is relatively short with respect to sequence length[2], the system does not have enough time to recover from bad/false tracks to detect periodic motions.

To evaluate the effect of non-stationary periodic distractions such as walking and running actions, we report the false positive detection rate of our approach when applied to the walking and running subset of UCF-ARG and KTH datasets. Table III shows the results. The pipeline shows zero false detections on either "running" sequence. However it exhibits a 13% and 16.67% false detection rate on "walking" sequences of KTH and ARG datasets respectively. This is mainly due to tracking and motion estimation error which causes a non-stationary periodic object to appear stationary. The false detection rate can be reduced by decreasing δ_{max} at the expense of a lower detection rate or slower response time. An alternative is to use robot behavior to reject false positives as discussed in Sec. V.

[2]Average duration of waving gesture sequences in KTH and UCF-ARG datasets are 21.5 and 10.2 seconds respectively.

TABLE II: The accuracy of hand waving detection for each experiment (DT: Detection Rate, FDR: False Detection Rate, MR: Miss Rate)

Dataset	Number of Actions	DR	FDR	MR
KTH	100	78%	0%	22%
ARG	48	56.25%	4.16%	39.58%

TABLE III: False Detection Rate (FDR) for walking and running actions (UCF-ARG and KTH datasets)

Dataset	Action	Number of Actions	FDR
KTH	Walk	100	13%
KTH	Run	100	0%
ARG	Walk	48	16.67%
ARG	Run	48	0%

B. Closed-loop experiments with outdoor UAV

To demonstrate the effectiveness of our system in establishing joint attention between a flying robot and a human, we performed a series of 22 trials in outdoor settings. The trials were carried out on three different days, under two lighting conditions (sunny and overcast), at two different locations and with different subjects. Both locations were open grass fields with trees and bushes at one side. In each trial the robot traversed a pre-defined path (a set of GPS waypoints) of length 10 meters back and forth at a fixed altitude and heading. We designed the UAV's flight path such that the vegetation be visible at all times. The altitude was varied from 10 meters to 15 meters during the trials. In each trial, a single human tries to grab the UAV's attention by waving at it. Two types of distractions were present in the field of view of the UAV: walking, running or standing people and natural distractions such as trees and bushes often moving in the wind. Since the robot does not perform any active searching to find humans, the workspace in which the subject and human distractors are allowed to act is marked in advance. The robot is fully autonomous, untethered and self-contained except during take-off and landing, where it is controlled by a human safety pilot. The script for each trial is as follows:

- Human distractors perform their act during the entire length of a trial and are instructed to stay with the UAV's workspace
- The human subject chooses an arbitrary position in the workspace prior to the start of the trial
- The UAV takes off and flies back and forth between two

Fig. 4: Example images from the robot's perspective during experiments. Location 1 (left) Location 2 (right)

predefined points
- The subject is instructed to stand still while the robot traverses the first leg (from A to B). This is to test that the system correctly handles the absence of gestures.
- Once the robot is on the return leg (after reaching point B) the subject starts waving
- If the UAV detects this gesture it stops translating, hovers, and starts the "wobble" behavior. This indicates to the waving human that she is detected. This successfully concludes a trial.
- If the UAV reaches a waypoint without detecting a waving gesture it starts a new traverse back to the previous location. The subject is allowed to try again to get the robot's attention. We report the number of retries in the results section. Runs with more than 1 retries are considered failures.

For a few trials we asked the subject not to try to attract the UAV's attention so we can examine the system's resistance to false positives. Figure 4 shows the robot's field of view during trials on two different locations. Table IV summarizes the conditions and results of all trials. The overall success rate of all trials was 81.8%. During all 22 trials, the UAV was never attracted to a false positive. In 5 successful runs with a waving human subject, it took the UAV one more traversal to find and acknowledge the subject.

Analyzing the experimental data we observe two major causes of failures. In two trials the human was on the edge or out of robot's field of view for the majority of time. Therefore tracking the human was not reliable enough to detect periodicity. This was mainly due to errors in the UAV's waypoint navigation and position control which changed the robot's visible workspace. Since the robot is flying several meters away from the human and the camera is barely visible, the subject was not able to estimate the robot's field of view to correct her location. This emphasizes the importance of the robot providing behavioral feedback when the human is detected. In two other failed trials, the vision system was not able to detect a moving object. Either too few features were detected on the subject's body or they were too sparse to form a cluster. We provide some suggestions to mitigate these problems in Sec. V.

C. Runtime Performance

For all three experiments, we measured the execution time per frame incurred by each step of the vision pipeline. The last column of Table I shows the average processing time per frame for each experiment. Table V shows the detailed breakdown of execution time for each component of the pipeline during each experiment. The processing time is less than the inter-frame time, so the system works in real-time.

TABLE IV: Outcome of all trials. C1: Location 1, Late Afternoon, Overcast, C2: Location 1, Noon, Overcast, C3: Location 2, Noon, Overcast, C4: Location 2, Late Afternoon, Sunny, (r): running, (w): walkig

Trial	Condition	Subjects and Distractors	Alt.	# of Tries	Outcome
1	C1	1,0	10	1	Success
2	C1	1,1(w)	10	1	Success
3	C1	0,1(w)	10	N/A	Success
4	C2	0,2(r)	10	N/A	Success
5	C2	1,1(w)	12	2	Success
6	C2	1,1(w)	12	4	Failure
7	C2	1,2(w)	12	3	Failure
8	C2	0,1(w)	12	N/A	Success
9	C2	1,0	12	1	Success
10	C3	1,1(w)	12	2	Success
11	C3	1,1(w)	12	4	Failure
12	C3	1,1(w)	12	1	Success
13	C3	0,3(w)	12	N/A	Success
14	C3	1,1(r)	15	1	Success
15	C3	1,1(r)	15	2	Success
16	C3	1,1(r)	15	1	Success
17	C3	1,1(r)	15	2	Success
18	C4	1,2(w)	15	4	Failure
19	C4	1,1(r)	15	1	Success
20	C4	1,1(r)	15	2	Success
21	C4	1,1(r)	15	1	Success
22	C4	1,1(w)	15	2	Success
Overall Success Rate					81.8%

TABLE V: Mean per-frame execution time breakdown for each component of the pipeline (in milliseconds).

	KTH	ARG	UAV
Pre-processing	0.179	2.14	2.51
Feature Detection & Tracking	0.529	13.10	13.69
Find Homography	2.30	10.55	10.10
Salient Object Detection	0.08	1.89	5.25
Object Tracking & Periodicity Detection	0.052	0.17	0.10
Total	**3.13**	**27.84**	**31.65**
Stddev	1.51	3.22	2.86

V. Conclusion and Future Work

In this paper we presented the first demonstration of human-UAV interaction in outdoor environments using real-time computer vision running entirely on-board. We show how a dual arm-waving gesture can be used to attract a flying robot's attention while being robust to similar distractions such as walking and running people. By acknowledging the user through a wing wiggle, the robot communicates its readiness for further interaction with the user.

The main limitation of the current approach is that the UAV can become attracted to non-interesting stationary periodic motions caused either by other human actions (e.g. digging) or irrelevant extrinsic processes (e.g waving trees). In future work we will explore two approaches to overcome this limitation. The first approach is to use robot behavior to reject false positives e.g. by approaching the target and

performing close-range inspection/interaction. The second approach is to use more discriminative motion features and classification techniques to detect the gesture.

As discussed in Sec. IV-B, it is not trivial for the user to estimate robot's field of view from distance, thus the user may not always be able to place herself in the robot's FOV to grab its attention. We are planning to study two possible solutions. The first approach is based on sound, visual or behavioral feedback from the UAV to the human during the pre-interaction phase. This is to help the user to better understand the UAV's intention and internal state. The second possible approach is to explore the area at high altitude (and thus larger camera footprint) and adaptively focus the UAV's active search on areas with interesting motion features [37]. Thus, the UAV can explore regions with high probability of human presence in detail and with smaller chance of missing her. These approaches are complementary: both promise to increase the ability we demonstrated in this paper of a UAV in flight to detect, track and establish shared attention with a human user.

ACKNOWLEDGMENTS

Support by the NSERC Canadian Field Robotics Network.

REFERENCES

[1] 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 Int. Conf. on*, Nov 2013, pp. 617–623.

[2] M. Lichtenstern, M. Frassl, B. Perun, and M. Angermann, "A prototyping environment for interaction between a human and a robotic multi-agent system," in *Proc. of the Int. Conf. on Human-Robot Interaction, (HRI)*, 2012, pp. 185–186.

[3] T. Naseer, J. Sturm, and D. Cremers, "FollowMe: Person following and gesture recognition with a quadrocopter," in *2013 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2013, pp. 624–630.

[4] G. Costante, E. Bellocchio, P. Valigi, and E. Ricci, "Personalizing vision-based gestural interfaces for HRI with UAVs: a transfer learning approach," in *2014 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2014)*. IEEE, 2014, pp. 3319–3326.

[5] "Wilderness Survival Merit Badge Handbook," Boy Scouts of America, 1998.

[6] S. Pourmehr, V. Monajjemi, J. Wawerla, R. T. Vaughan, and G. Mori, "A robust integrated system for selecting and commanding multiple mobile robots." in *Robotics and Automation ICRA, IEEE Int. Conf. on*. IEEE, 2013, pp. 2874–2879.

[7] T. Mantecón, C. R. del Blanco, F. Jaureguizar, and N. García, " New generation of human machine interfaces for controlling UAV through depth-based gesture recognition," *SPIE Defense + Security*, vol. 9084, pp. 90 840C–90 840C–11, Jun. 2014.

[8] D. Perez, I. Maza, F. Caballero, D. Scarlatti, E. Casado, and A. Ollero, "A Ground Control Station for a Multi-UAV Surveillance System," *J. of Intelligent and Robotic Systems*, vol. 69, no. 1-4, pp. 119–130, 2013.

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[10] J. Sokalski, T. P. Breckon, and I. Cowling, "Automatic salient object detection in uav imagery," in *Proc. th Int. Conf. on Unmanned Air Vehicle Systems*. Proc. 25th Int. Conf. on . . . , 2010, pp. 11.1–11.12.

[11] G. R. Rodríguez, S. Thomas, J. del Cerro, A. Barrientos, and B. MacDonald, "A Real-Time Method to Detect and Track Moving Objects (DATMO) from Unmanned Aerial Vehicles (UAVs) Using a Single Camera," *Remote Sensing*, vol. 4, no. 4, pp. 1090–1111, Apr. 2012.

[12] M. Siam and M. ElHelw, "Robust autonomous visual detection and tracking of moving targets in UAV imagery," vol. 2, pp. 1060–1066.

[13] M. Kimura, R. Shibasaki, X. Shao, and M. Nagai, "Automatic extraction of moving objects from UAV-borne monocular images using multi-view geometric constraints ," in *Int. Micro Air Vehicles*, 2014.

[14] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, pp. 1–43, 2011.

[15] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Computer Vision*. IEEE, 2003, pp. 726–733 vol.2.

[16] C.-C. Chen and J. K. Aggarwal, "Recognizing human action from a far field of view," pp. 119–125, 2009.

[17] S. Wu, O. Oreifej, and M. Shah, "Action recognition in videos acquired by a moving camera using motion decomposition of Lagrangian particle trajectories," in *2011 IEEE Int. Conf. on Computer Vision (ICCV)*. IEEE, 2011, pp. 1419–1426.

[18] M. Allmen and C. R. Dyer, "Cyclic motion detection using spatiotemporal surfaces and curves," in *Pattern Recognition, 1990. Proc., 10th Int. Conf. on*. IEEE Comput. Soc. Press, 1990, pp. 365–370.

[19] 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, Dec. 1994.

[20] F. Liu and R. W. Picard, "Finding periodicity in space and time," in *Computer Vision, 1998. Sixth Int. Conf. on*, 1998, pp. 376–383.

[21] Y. Ran, I. Weiss, Q. Zheng, and L. S. Davis, "Pedestrian Detection via Periodic Motion Analysis," *Int. J. of Computer Vision*, vol. 71, no. 2, pp. 143–160, Feb. 2007.

[22] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 22, no. 8, pp. 781–796, 2000.

[23] R. Polana and R. C. Nelson, "Detection and Recognition of Periodic, Nonrigid Motion," *Int. J. of Comp. Vision*, vol. 23, pp. 261–282, 1997.

[24] P. V. K. Borges, "Pedestrian Detection Based on Blob Motion Statistics," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. 23, no. 2, pp. 224–235, 2013.

[25] X. Tong, L. Duan, C. Xu, Q. Tian, H. Lu, J. Wang, and J. S. Jin, *Periodicity Detection of Local Motion*, 2005.

[26] J. Sattar and G. Dudek, "Where is your dive buddy: tracking humans underwater using spatio-temporal features," in *Intelligent Robots and Systems*. IEEE, 2007, pp. 3654–3659.

[27] A. Ollero, J. Ferruz, F. Caballero, S. Hurtado, and L. Merino, "Motion compensation and object detection for autonomous helicopter visual navigation in the COMETS system," in *Robotics and Automation, 2004. Proc. ICRA. 2004 IEEE Int. Conf. on*. IEEE, 2004, pp. 19–24.

[28] F. Caballero, L. Merino, J. Ferruz, and A. Ollero, "Vision-Based Odometry and SLAM for Medium and High Altitude Flying UAVs," in *Unmanned Aircraft Systems*, 2008, pp. 137–161.

[29] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Euro. Conf. on Comp. Vision*, vol. 1, 2006, pp. 430–443.

[30] J.-Y. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker: Description of the Algorithm," Intel Corporation Microprocessor Research Labs, Tech. Rep., 2000.

[31] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Proc. of the Second Int. Conf. on Knowledge Discovery and Data Mining (KDD-96)*, 1996, p. 226231.

[32] F. Luetteke, X. Zhang, and J. Franke, "Implementation of the hungarian method for object tracking on a camera monitored transportation system," in *Proc. of 7th German Conf. on Robotics*, 2012, pp. 1–6.

[33] T. Aittokallio, M. Gyllenberg, O. Nevalainen, and O. Polo, "Testing for periodicity in signals: An application to detect partial upper airway obstruction during sleep." *J. of Theoretical Medicine*, vol. 3, no. 4, pp. 231–245, 2001.

[34] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly, 2008.

[35] I. Laptev and B. Caputo. (2004) Recognition of human actions. [Online]. Available: http://www.nada.kth.se/cvap/actions/

[36] (2008) University of Central Florida, UCF aerial camera, rooftop camera and ground camera dataset. [Online]. Available: http://vision.eecs.ucf.edu/data/UCF-ARG.html

[37] S. A. Sadat, J. Wawerla, and R. Vaughan, "Fractal trajectories for online non-uniform aerial coverage," in *Robotics and Automation (ICRA), 2015 IEEE Int. Conf. on*, 2015.