Gesture-Based Control of Autonomous UAVs

(Extended Abstract)

Jon Bolin[*]
jonathon-bolin@utulsa.edu

Chad Crawford[*]
chad-crawford@utulsa.edu

William Macke[*]
william-macke@utulsa.edu

Jon Hoffman[*]
jon-hoffman@utulsa.edu

Sam Beckmann[*]
sam-beckmann@utulsa.edu

Sandip Sen[*]
sandip-sen@utulsa.edu

ABSTRACT

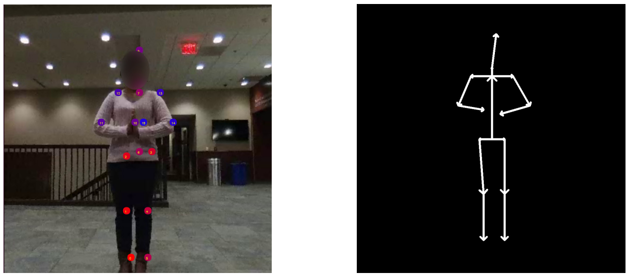
Unmanned Aerial Vehicles (UAVs) have been traditionally controlled via remote control or by software, which require skill using the remote or expert programming skills. Our goal is to develop a natural mode of directing a drone's actions, akin to the forms of expression one finds between a person and a pet and hence accessible to almost any person without specialized training or expertise in using electronic gadgets. We build on prior work on analyzing video streams to use the video from the drone's on-board camera to enable gesture-based control. Our approach uses a pre-trained convolutional neural network for pose extraction, Haar cascades to identify regions of interest within the UAV's field of view, and a finite state machine to select the drone's action.

CCS Concepts

•**Human-centered computing** → **Collaborative interaction; Gestural input;** *Human computer interaction (HCI); Interface design prototyping;*

Keywords

HCI, Gesture Recognition, Human-UAV Interaction



(a) Joints located by convolutional neural network.

(b) Vector representation of the identified joints.

Figure 1: Point and vector form of home position.

1. INTRODUCTION

Existing methods for controlling UAVs have primarily relied on two modalities for controlling the actions of a UAV: (a) use of remote controls, or (b) use of software programs to pre-configure the flight paths, movements, and other actions. Our premise is that humans are adept and efficient in using physical gestures to communicate, particularly for directional signals. There is a rich body of work in image analysis and human-computer interaction on gesture recognition as well as on intent inference from observed movements. To the best of our knowledge, however, there has been little work to develop gesture-based control of autonomous UAVs, and even less using the on-board camera. We posit that gesture-based communication can be used as the most effective UAV control mechanism by non-expert users in a number of UAV applications, particularly those involving Human-UAV (HUAV) collaboration.

The domain of Human-Robot Interaction is a widely researched topic with a variety of applications [4, 5]. The growth in popularity of RGB plus Depth (RGB-D) sensors has made gesture and pose recognition a simpler task for robots equipped with such sensors [6, 7, 12]. In several works, researchers demonstrated this potential for gesture-based control when aided by these RGB-D sensors [10, 11]. However, UAVs are not equipped with these sensors, making gesture recognition a more complex task from the on-board camera [2, 8].

2. EXPERIMENTAL FRAMEWORK

In order to extract a human pose from an RGB image, we employ a two-step process that (1) identifies the locations of 16 key points defined by the MPII Human Pose Database [1] then (2) maps the identified points to one of a set of predefined poses that the UAV will interpret as a command.

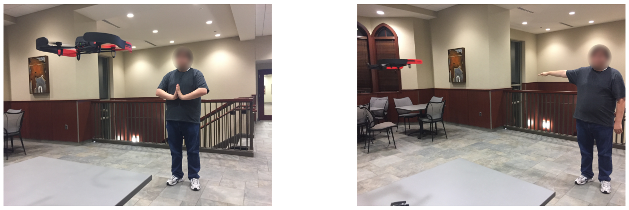
2.1 Pose Identification

To keep the poses lightweight, we convert 8 of the points from the output of the stacked hourglass model by Newell et al. [9] into 5 normalized vectors, and use 4 of the dot products of those vectors to classify the pose. The vector and point forms are illustrated in Figure 1.

Similarly to how a pet has a name that can be used to get its attention, the UAV needs a way to distinguish be-

[*] Tandy School of Computer Science, The University of Tulsa, Oklahoma, USA

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*



(a) Before the gesture to move.

(b) UAV moves following the gesture.

Figure 2: HUAV collaboration: following gesture for movement.

tween stray arm movements and command gestures. All executable commands must be preceded by the Home position to get the UAV's attention. If the UAV is not flying, it uses the Home position as a cue to take off.

2.2 Action Selection

With the pose identified, the UAV must decide on an appropriate action to take. The maneuver the UAV undertakes is determined by its current state and the most recently identified pose. Action selection uses a finite state machine to determine the action to be performed. If the UAV is not listening, it will only move on to a different state if it identifies the home position. Then, it will listen to the next pose to determine the appropriate action.

3. RESULTS

We were able to successfully direct the UAV using only upper-body gestures, but not to the extent of natural person-pet interaction. Our pose recognition mechanism is evaluated for detection accuracy and response time.

3.1 Upper-Body Control of UAV

The sequence of events for directing the UAV's actions, from taking off to landing, is as follows. First, the UAV will take off when it identifies a person in the home position. Once the UAV is in the air, it follows the person with its camera by physically rotating for left and right movement and by adjusting the camera's direction for vertical movement. While doing this, it listens for the home position, and, if the home position is identified, it will listen for the command. At this point, it will execute the command until the command gesture is discontinued or it loses track of the operator, at which point it will "unlock" and listen for the home position again. This continues until the land command is given, at which point the UAV will stop flying and the program will exit.

One such example of a gesture-based control is given in Figure 2. The home position tells the UAV to listen to the human's command, and Figure 2(b) shows the resulting leftward movement by the UAV.

There are presently ten categories of motion for which gestures have been defined: takeoff, land, up, down, left, right, forward, backward, transfer control left, and transfer control right. Some command poses were intuitive, such as left and right being simply pointing in the direction the UAV should move, whereas others were less inherent in nature due to the present limitations in our gesture recognition method, namely the "come closer" and "go further away" commands.

Gesture Duration (s)	False Positive Rate	False Negative Rate	Misinterpreted Gestures	Successful Gestures
1	0.03333	0.16667	0.03333	0.8333
2	0.01667	0.06667	0.01667	0.9833
3	0.05	0.06667	0	0.9833
4	0.06667	0.18333	0.01667	0.9833
5	0.01667	0.26667	0	0.9833

Figure 3: Gesture Recognition Performance

While the natural gesture one may expect would be waving for the UAV to come near or shooing it away, our gesture recognition is limited to subsequent pose identifications at time intervals exceeding 0.5 seconds, so we developed a slightly less natural method of communicating these intents.

3.2 Evaluation

In order to evaluate the false positive and negative rates of the pose detection, we recorded a series of gestures from the UAV's video stream, manually identified the correct pose classification, and passed the stream through the UAV's video processing algorithms to identify the UAV's classification of the gestures. The gestures were held for a specified interval ranging from 1-5 seconds. We ran two trials with each trial consisting of 30 gestures. Figure 3 displays the false positive and false negative rates for all five tests aggregated across the two trials.

The false positive rate is the proportion of gestures that the UAV identified but that were not performed by the operator. The false negative rate is each frame that the UAV failed to identify a gesture despite the operator having performed it, which includes gestures the UAV lost track of and gestures that the UAV missed completely. This was strongly influenced by the noise from the neural network, which caused the UAV to lose track of a gesture while it was still being performed. As such, its value tends to get higher as the duration of the gesture increases. Misinterpreted gestures are those that the UAV misidentified as being a different pose than was actually performed, and successful gestures represents the proportion of total gestures that were correctly identified.

To evaluate the UAV's response time, we used the same video streams and calculated the delay between the manual identification and the UAV's recognition. On average, it takes roughly 0.8 seconds to identify the pose.

Discussion & Conclusions:.

Our current method of pose detection is effective but slow, and the best way to improve would be to either decrease its run time or employ some form of an motion-tracking algorithm to get an approximation of the locations in between the outputs from the network. Lucas-Kanade optical flow [3] would allow for tracking the joints as the person moves.

We currently have control of the UAV in all three coordinate axes based solely on gestures, and its rotation can be controlled by the position of the human controlling it as well as by passing control to another person. As this project continues to develop, we will attempt to allow for movement-based gestures and to improve on the current framework.

REFERENCES

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693. IEEE, 2014.

[2] P. Barros, G. I. Parisi, D. Jirak, and S. Wermter. Real-time gesture recognition using a humanoid robot with a deep neural architecture. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 646–651. IEEE, 2014.

[3] J.-Y. Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 5(1-10):4, 2001.

[4] X. Chen and M. Koskela. Online rgb-d gesture recognition with extreme learning machines. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 467–474. ACM, 2013.

[5] M. A. Goodrich and A. C. Schultz. Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, 1(3):203–275, 2007.

[6] J. Han, L. Shao, D. Xu, and J. Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE transactions on cybernetics*, 43(5):1318–1334, 2013.

[7] K. Khoshelham and S. O. Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012.

[8] J. Nagi, A. Giusti, G. A. Di Caro, and L. M. Gambardella. Human control of uavs using face pose estimates and hand gestures. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 252–253. ACM, 2014.

[9] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[10] K. Pfeil, S. L. Koh, and J. LaViola. Exploring 3d gesture metaphors for interaction with unmanned aerial vehicles. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 257–266. ACM, 2013.

[11] A. Ramey, V. González-Pacheco, and M. A. Salichs. Integration of a low-cost rgb-d sensor in a social robot for gesture recognition. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 229–230. ACM, 2011.

[12] Y. Yao and Y. Fu. Real-time hand pose estimation from rgb-d sensor. In *2012 IEEE International Conference on Multimedia and Expo*, pages 705–710. IEEE, 2012.