Human Control of UAVs using Face Pose Estimates and Hand Gestures

Jawad Nagi, Alessandro Giusti, Gianni A. Di Caro, Luca M. Gambardella
Dalle Molle Institute for Artificial Intelligence (IDSIA), CH-6928 Lugano, Switzerland
{jawad,alessandrog,gianni,luca}@idsia.ch

ABSTRACT

As a first step towards human and multiple-UAV interaction, we present a novel method for humans to interact with airborne UAVs using locally on-board video cameras. Using machine vision techniques, our approach enables human operators to command and control Parrot drones by giving them directions to move, using simple hand gestures. When a direction to move is given, the robot controller estimates the angle and distance to move with the help of a *face score system* and the *estimated hand direction*. This approach offers mobile robots the ability localize with human operators and provides UAVs/UGVs with a better perception of the environment around the human.

Categories and Subject Descriptors

I.2.9 [**Robotics**]; I.4 [**Image Processing and Computer Vision**]; I.5 [**Pattern Recognition**]: General

Keywords

Human-robot interaction, face pose, gesture recognition

1. INTRODUCTION

Hand gestures are easily recognizable and have the advantage of being easy to use, as well as being natural and intuitive means for human-robot interaction (HRI). For humans to command and control UAVs/UGVs using hand gestures, robots need to know where the human is located (or positioned) in the environment. With the recent advances in computational power and machine vision algorithms, the detection of humans has become relatively easy [1].

In order for a human to be able to command a robot to move a specific position, based on the direction of the human hand, a robot needs to know the orientation of the hand direction with respect to the *face pose* of the human based on its field of view. At the aim of visual interaction between humans and robots, in this work we adopt *face detection* to identify if a human is present in the field of view of the UAV. We consider that interaction only initiates after a human face is detected by the robot [2]. If a face is detected by the robot, then we consider the pose of the face as the *angle* between the human and the robot's point of view. Since, a robot also needs to take into account the *proximity* (distance) between itself and the human when maneuvering, we estimate this distance using measures of the detected face.

To estimate the direction (given by a hand gesture) where the robot should move to, we compute the orientation of the *hand direction* with respect to the location of the face. Using measures of the estimated face pose together with the proximity measure and hand direction, a UAV can estimate its current position (angle,distance) in the environment with respect to the human, and can move along a path to the direction specified by the human.

2. METHODS

2.1 Face Score System

To detect and track a human face, we use the built-in video camera of the Parrot that acquires images in resolution of 1280×720 pixels at 30 fps. Acquired frames (images) are wirelessly transmitted from the Parrot onto a Linux machine for offline processing. We perform face detection using the OpenCV implementation of the Viola-Jones face detector. Given images such as in Figure 1, we locate rectangular areas in images that contain a human face from different poses (angles), e.g. a frontal poses (see Figure 1(b)) or side poses (see Figures 1(a) and (c)). After a face is detected, we employ a Kalman Filter for tracking the detected face and also for reducing false positive detections. As face detectors are insensitive to small changes in scale or position, multiple sub-windows are typically clustered around the face.

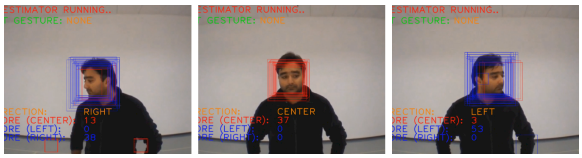

Figure 1: Face pose estimation using a flying UAV. Identified face poses: (a) right, (b) center, (c) left.

The Viola-Jones face detector, a cascade Haar classifier, is used for finding groups of neighbouring sub-windows around the human face. In this work, we employ two pretrained Haar classifiers to detect faces from multiple point of views. One classifier detects frontal poses of the face profile C_{front},

HRI'14, March 3–6, 2014, Bielefeld, Germany.
ACM 978-1-4503-2658-2/14/03.
http://dx.doi.org/10.1145/2559636.2559833.

and the other detects side profiles of the face C_{side}. As two classifiers are used, the number of red-colored sub-windows in Figure 1 indicate the face detections from C_{front} (center), whereas the blue-colored sub-windows show detections from the C_{side} (left and right). Next, we determine the *number of neighbouring sub-windows* detected from the frontal and side face poses for estimating the relative pose (position) of the face with respect to the robots point of view. For every frame (image) acquired by the robot, 4 face quality measures are computed using C_{front} and C_{side}:

1. Classifier C_{front} is run on frame f to obtain $score_{front}$.
2. Frame f is flipped horizontally (180°) with subsequent horizontal shift to obtain, f_{hor}. Classifier C_{side} is run on f_{hor} to obtain $score_{frontflip}$.
3. Classifier C_{side} is run on frame f to obtain $score_{side}$.
4. Using frame f_{hor} (obtained in Step 2), and running classifier C_{side} results in $score_{sideflip}$.

The current direction of the human face (i.e., left, center and right) with respect to the robot's point of view, is computed using:

$$S_{center} = score_{front} + score_{frontflip} \quad (1)$$

$$S_{right} = score_{side} \quad (2)$$

$$S_{left} = score_{sideflip} \quad (3)$$

where equations (1), (2) and (3) each represent independent *face scores* assessing the quality of the detected face from a 3-dimensional perspective. In simpler words, when the human is directly looking towards the robot, the value of S_{center} is higher than S_{right} or S_{left}. Similarly, if the human's face is directed towards the left, then S_{left} is higher than S_{right} and S_{center}, and vice versa when the human is looking towards the right.

We estimate the relative distance d between a human and a UAV by using information from the face detector (i.e., the *average size of all detected sub-windows* around the face). Using the computed face measure information $X = \{S_{left}, S_{center}, S_{right}, d\}$ with known ground truth (i.e., *known face pose* Y; value in between closed interval $[0, 180°]$), we adopt the Support Vector Regression (SVR) with a non-linear kernel function to learn the mapping (relationship) between X and Y. For a set of X input features, the SVR predicts the *face pose* θ onto a horizontal plane of $[0, 180°]$ using a pre-trained SVR, where 0°, 90° and 180° represent the left, center and right respectively (see Figure 1).

2.2 Hand Direction

Directions for the robot to move are provided by the human operator using the left or right hand, as illustrated in Figure 2. For simplicity purposes, we use hand gestures naturally supplemented by fine controls through tangible input gadgets (i.e., coloured gloves). By pointing the hand towards the left or right, the human can command the robot to move towards a specific direction.

For estimating the direction of the hand with respect to θ for frame f, the multiple sub-windows resulting from the multiple face detectors are averaged together to form a single rectangular area f_{rect} bounding the face, as illustrated by the face bounding boxes in Figure 2. Next, the centroid of f_{rect} is computed as fc, the *face centroid*. Performing color-based segmentation followed by blob analysis on the coloured glove (hand), the *hand centroid* hc can be easily computed. The *hand orientation* H_{dir} (covering a circular area of $\pm 180°$ degrees) with respect to fc can then be computed using, $H_{dir} = (\text{atan2}(hc - fc) * 180/\pi)$.

2.3 Robot Control

At every control step t, a frame f acquired by the robot is used to compute, $X^t = \{S^t_{left}, S^t_{center}, S^t_{right}, d^t\}$ (a feature vector), where d_t represents the average area of the face sub-window. While moving to a position commanded by the human, the robot makes use of the feedback information (θ^t, d^t) from the face score system at every t. To reach the directed position H_{dir}, the robot controller aims to 'minimize' the relative angle between the current position of the UAV and the target destination: $\min\{P^t_{target}\}$, where $P^t_{target} = abs(\theta^t - H_{dir})$ (absolute difference to move in degrees), while maintaining the distance d^t. When P^t_{target} goes lower than a predefined threshold, the robot stops, as it reaches the position commanded by the human.

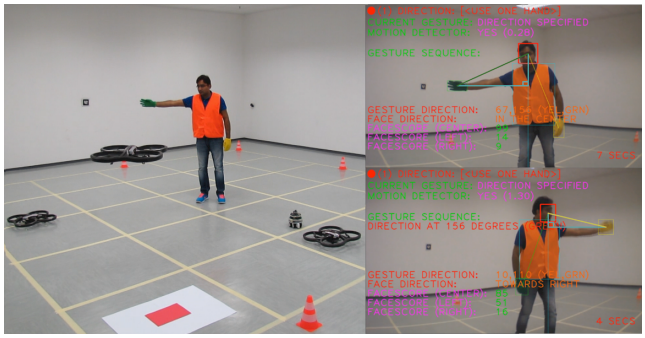

Figure 2: A human operator commanding a flying UAV to move in specific directions.

3. CONCLUSION

We presented a machine vision approach directly applicable for human-UAV localization problems. Our approach makes use of human faces poses in order to allow UAVs to estimate their position (location) in the environment with respect to the human. To determine the angle and distance where the robot should move, the robot controller makes use of the feedback information from the face score system together with the estimated hand direction.

4. ACKNOWLEDGMENTS

This research is supported by the Swiss National Science Foundation (SNSF) through the National Centre of Competence in Research (NCCR) Robotics.

5. REFERENCES

[1] B. Milligan, G. Mori, and R. T. Vaughan. Selecting and commanding groups of robots in a vision based multi-robot system. Proc. of the 6th ACM/IEEE Intl. Conf. on HRI (Video Session), 2011.

[2] S. Pourmehr, M. Monajjemi, R. T. Vaughan, et al. You two! take off!: Creating, modifying and commanding groups of robots using face engagement and indirect speech in voice commands. In *Proc. of the IEEE International Conference on IROS*, 2013.