Monocular Vision-based Human Following on Miniature Robotic Blimp

Ningshi Yao, Emily Anaya, Qiuyang Tao, Sungjin Cho, Hongrui Zheng and Fumin Zhang

***Abstract*— We present an approach that allows the Georgia Tech Miniature Autonomous Blimp (GT-MAB) to detect and follow a human. This accomplishment is the first Human Robot Interaction (HRI) demonstration between an uninstrumented human and a robotic blimp. GT-MAB is an ideal platform for HRI missions because it is safe to humans and can support sufficient flight time for HRI experiments. However, due to complex aerodynamic influence on the blimp, the human following task for GT-MAB with a single on-board camera is a challenging problem. We integrate Haar face detector and KLT feature tracker to achieve robust human tracking. After a human face is detected in the real-time video stream, we estimated the 3D positions of the human with respect to GT-MAB. Vision-based PID controllers are designed based on estimated relative position and the motion primitives of GT-MAB such that it can achieve stable and continuous human following behavior. Experimental results are presented to demonstrate the human following capability on GT-MAB.**

MULTIMEDIA MATERIAL

A video attachment to this work is available at: https://www.youtube.com/watch?v=cp1201phPts.

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) can be used for many applications. Detection of dynamic objects using flying robots can have a profound impact on applications such as traffic supervision, autonomous robot navigation and surveillance of large facilities. These environments require UAVs to cooperate with humans. However, typical UAVs such as quad-rotor helicopters and aircrafts can be dangerous when interacting with humans due to their sharp and powerful propellers. In addition, their flight time is relatively short. A safer robot that can fly for longer time would be beneficial.

We developed the Georgia Tech Miniature Autonomous Blimp (GT-MAB) as an alternative UAV for indoor experiments that can support research on 3D motion control and human-robot interaction [1]. GT-MAB does not hurt human even when it accidentally collides with humans because it is very light. As illustrated in figure 1, a human can stand very close to GT-MAB and will not be scared. In addition to being safe, GT-MAB has a relatively long flight time of over 2 hours per battery charge. In this paper, we achieve

This work is partially supported by ONR grants N00014-14-1-0635 and N00014-16-1-2667; and NSF grants IIS-1319874 and CMMI-1436284.

Ningshi Yao, Qiuyang Tao, Sungjin Cho and Fumin Zhang are with the School of Electrical and Computer Engineering, Georgia Institute of Techonology, Atlanta, GA 30308, USA. Email: {nyao6, qtao7, scho88, fumin}@gatech.edu

Emily Anaya is with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA. Email: {eanaya}@wisc.edu

Hongrui Zheng is with the College of Computing, Georgia Institute of Techonology, Atlanta, GA 30308, USA. Email: {hzheng40}@gatech.edu

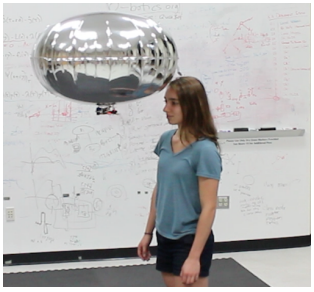


Fig. 1. An uninstrumented person interacts closely with GT-MAB

the human following behavior on GT-MAB. To the best of our knowledge, this is the first HRI demonstration between an uninstrumented human and an autonomous robotic blimp using only passive computer vision. This accomplishment is an important step towards safer human robot interaction capabilities such as dancing with humans, serving as a flying telepresence robot, leading people to a destination and assisting workers in industry. The unique physical design of GT-MAB enables the use of one on-board monocular camera as the only sensor to achieve human detection and human position estimation.

Plenty of work exists for UAVs and human interaction. Interfaces, such as touch screen [2], audio/speech [3] and computer vision [4], [5], are applied to achieve HRI missions for flying vehicles. Vision-based human feature detection, such as face detection and gesture detection, are natural communication cues for robots to interact with uninstrumented humans who do not wear additional tracking devices [6]. People install cameras on the UAVs and use image processing techniques to enable the vehicles to sense the environment and recognize the commands from humans. A human controlling a team of quad-rotors through face and gesture recognition is presented in [7]. The quad-rotors are able to take pre-defined actions, such as joining the team, leaving the team or landing, based on human's gestures. In [8], Naseer et. al presented an approach to achieve the behavior of a single UAV following a single human using an on-board camera. However, a depth camera is required for this approach instead of a traditional RGB camera. Human following behavior based on monocular vision is presented in [9], requiring that the initial height of the UAV is known. A commercial product named "hover camera" can achieve static hovering and human following functionalities to take pictures and videos of a human, but it requires an extra down-looking camera and a sonar to stabilize the UAV [10]. Besides, it can only support 8 minutes of flight time.

Such vision-based human or target following missions are

challenging for quad-rotor helicopters or fixed-wing UAVs because of the flying postures and vibration of the platform. Normally, gimbal systems are used to adjust the position of on-board camera and rubber balls are used to absorb the vibration [11].

In contrast to other UAVs, robotic blimps are self-stabilized and have less vibration when flying, which are better platforms to use an on-board camera as the vision sensor. Besides, from the human users' point of view, blimps are better for HRI missions, taking social factors such as noise and appearance issues into consideration [12]. In [13], authors developed a spherical robotic blimp. Computer vision algorithms are developed to monitor activities in an area patrolled by the blimp. However, the spherical blimp is not able to distinguish a human from other objects and therefore, does not solely react to human motion.

In this work, we explore the possibility of detecting and following a human user with robotic blimp GT-MAB. We propose a design that follows a data pipeline: Detecting-Estimating-Following. We first implement the robust face detection and KLT feature point tracking algorithms to track the human face in the live video stream using the blimp camera. Then based on the face position in the 2D image frame, the relative 3D position (i.e., the relative orientation and distance between the blimp and human target) of the human can be estimated. The estimation method does not require calibration of the camera. The flight control for the blimp to follow the human is difficult due to the complex nonlinear dynamics of the blimp and the external aerodynamical influence. In our work, we utilize the motion primitives of GT-MAB and design PID controllers to achieve stable and continuous human following behavior on GT-MAB. We also design a side-way motion controller to regulate the blimp flight to maintain the human in sight of the camera.

This paper is organized as follows. In Section II, we introduce the hardware and system setup for the blimp and human interaction. In Section III, we present the method of localizing a human using human face detection and the control of the blimp. In Section IV, we show experimental results of the human following behavior implemented on GT-MAB. Section V is the conclusion.

II. GT-MAB Platform and Hardwares

GT-MAB consists of an envelope and a customized gondola. The envelope has a unique saucer-like shape, which solves the conflict between maneuverability and stability while maximizing the buoyancy with limited footprint. In addition, the Helium-filled envelope provides a lift force for the entire vehicle, which significantly extends the flight duration of the blimp since no energy is required to keep it aloft. The gondola is a 3D-printed mechanical structure accommodating all on-board devices underneath the envelope. Figure 2 demonstrates the structure of the gondola and indicates the main components installed on it. We use five motors for this experiment. The vertically mounted motors are used to change the altitude while the horizontal ones enable the blimp to fly horizontally and change the heading

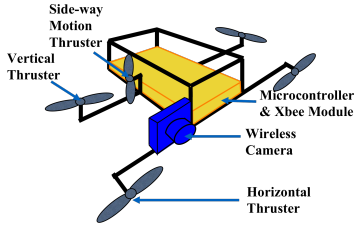


Fig. 2. The blimp gondola with various components located

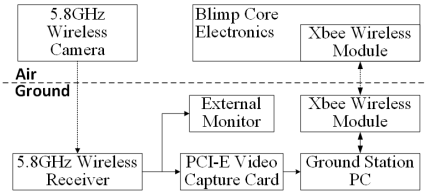


Fig. 3. System Overview

angle. One side-way motion motor is used to keep blimp in the front of human. The blimp has only 15 grams total load capacity, including the camera, microprocessors and wireless communication devices.

One difficulty for vision-based HRI mission on the robotic blimp is finding a camera that is light enough to be supported by the blimp. There is a trade-off between video quality and the weight of the camera. The camera we install on GT-MAB is a 5.8 GHz analog camera. This compact device weighs 4.5 grams and has diagonal field of view of 115 degrees. The camera is directly attached to the gondola. This camera is the best option we could find which can support wireless transmission. However, since the camera is analog, the video produced from it includes some glitch noise, which makes image processing more difficult than digital cameras.

Figure 3 shows the block diagram of the hardware setup for the system. Video stream coming from the on-board camera is obtained by the receiver and then digitized by the video capture card installed on the ground station PC. Outputs of the control algorithms running on the ground station PC are packed into commands with a certain format. Then the control command package is sent to the blimp via an Xbee wirelsss module.

III. Methodology

The implementation of human following behavior on GT-MAB involves three steps: 1) detecting a human face in real-time video stream, 2) estimating the relative position between the blimp and human, and 3) using vision-based estimation to control the movement of blimp to follow the human.

A. Human Detection and Tracking

Our work is based on human face detection because the face is the most distinctive feature separating a human from other objects. Many research works exist to improve the performance of human face detection algorithm [14], [15]. Viola and Jones were able to implement a robust face detection algorithm using Haar features and cascade classifier [16]. A Haar feature considers patterned adjacent rectangular regions at a specific location in a face image,

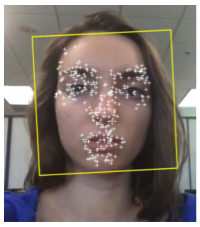


Fig. 4. A frame of the video stream from GT-MAB that was processed with face tracking algorithm

sums up the pixel intensities in each region and calculates the difference between these sums. The values computed based on Haar features are used to train a large number of weak classifiers whose detection qualities are slightly better than random guessing. Then weak classifiers are organized in classifier cascades using AdaBoost method [17] to form strong classifiers, which can robustly determine whether an image contains a human face.

We use two image sets for training Haar features, one is for front face and the other is for side face, so that the blimp can recognize a human from different angles. Due to the poor quality and noises from the blimp camera, the robust real-time face detection algorithm in [16] cannot guarantee continuous and reliable face detection in our case. To obtain stable detection of a human, rather than running human face detection every frame, our method uses the Kanade-Lucas-Tomasi (KLT) algorithm [18] to track the face after the human face is detected, which is computationally more efficient and robust than detecting the face each frame.

Algorithm 1 presents the pseudocode for the human face detection and tracking algorithm. The algorithm has two modes: face detection and KLT tracking. 1) The algorithm detects the human face using Haar features for the first several frames to prevent misdetection. Once a human face is detected, it extracts the feature points within the face region for the tracking mode. 2) In the face tracking mode, the algorithm matches the corner feature points of the new frame with the corner feature points from the previous frame, and it estimates the geometric displacement between these two sets of corner points. The displacement vector is applied to the previous face bounding box to obtain the new bounding box, so the algorithm can continuously track the human face. Once the number of corner points is below a certain threshold b, the mode switches back to face detection. A frame of the blimp video processed with algorithm 1 is shown in figure 4. The yellow rectangle is the bounding box that the algorithm recognizes as the area of the human face and the white crosses are the corner feature points.

Based on the bounding box of the human face, we can obtain the coordinates of the center of human face in image frame, denoted as $[i_P, j_P]^T \in \mathbb{R}^2$ and the face length l_f, where i_P, j_P and l_f are in units of pixels. And we use the variable S_f to record which side of the human is detected.

B. Relative Position Estimation

Our method localizes the blimp using vision from a camera only. This is different from most other blimps which utilize external localization system, suchs as GPS or indoor 3D localization.

Algorithm 1: Face Detection and Tracking

Data: Video Stream
Result: Face center $[i_P, j_P]^T$, face length l_f and side of the face S_f

$FrameNum = 1$, $FeatureNum = 0$;
while *Video is not ended* **do**
  if $FrameNum \leq a$ *or* $FeatureNum \leq b$ **then**
    /*Detection Mode*/
    Run frontal and side face detection and determine the bounding box of face;
    Save which side of face is detected in S_f;
    Detect corner points and re-initialize the KLT point tracker within the bounding box;
    Calculate the number of corner points $FeatureNum$;
  else
    /*Tracking Mode*/
    Estimate the geometric displacement between the corner points in previous frame and the corner points in the current frame;
    Apply the displacement vector to the bounding box in previous frame to obtain new bounding box;
    Run side face detection within the $\frac{1}{4}$ frame around the bounding box, set value for S_f if a side face is detected;
    Update $FeatureNum$ with the number of corner points within the new bounding box;
  Set the face center $[i_P, j_P]^T$ to be the center of bounding box and face length l_f to be the length of the bounding box;
  $FrameNum = FrameNum + 1$;

We assume that the camera satisfies the pinhole camera model [19], which defines the relationship between a 3D point $[x, y, z] \in \mathbb{R}^3$ in the camera coordinate $X_C - Y_C - Z_C$ and a 2D pixel $[i, j]^T$ in the image frame.

$$\begin{bmatrix} i \\ j \\ 1 \end{bmatrix} = \begin{bmatrix} f_i & 0 & i_0 & 0 \\ 0 & f_j & j_0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (1)$$

where f_i and f_j are the focal length in i and j directions, and $[i_0, j_0]^T$ is the optical center of the camera. Here we assume that f_x and f_y are both equal to the same focal length f and $[i_0, j_0]^T$ is the center of the image.

Reversely, if we know the focal length f of the camera and the actual depth z of the face center, we can reconstruct the 3D point $[x, y, z]^T$. However, since the wireless camera on GT-MAB is a monocular camera, we cannot directly obtain the accurate depth information of the human. In order to reconstruct the 3D point, we proposed a method to estimate the depth of the human face.

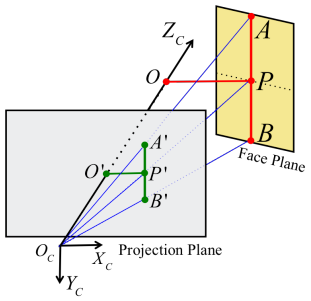


Fig. 5. Illustration of relative distance estimation

The illustration of human position estimation is shown in figure 5. Because the pitch and roll angles of blimp are very small, we can assume that the camera projection plane is always perpendicular to the ground, i.e. Y_C is perpendicular to the ground. This assumption does not hold for quad-rotors because quad-rotors need to change the pitch angle to fly forward or backward. Extra information is needed to estimate the height of quad-rotors [9], [20]. Blimp provides certain convenience to support vision-based HRI algorithms because the pitch and roll angles of the on-board camera can be controlled to stay at zero. Line AB represents the centerline of the human face and we assume it is parallel to the image plane, i.e. plane of human face is also perpendicular to the ground. Point $P = [x_P, y_P, z_P]^T$ is the center point of line AB. Points A', B' and P' are corresponding projection points. We denote the actual length of the human face as $L_0 := |AB|$ and denote the length of the human face in camera projection plane as $l_f := |A'B'|$.

First, we measure the human face length L_0 in units of meters. The human stands away from the camera at a fixed distance d_0 and the position of blimp is adjusted such that the center of the human face is at the center of image frame. Then we record how many pixels the human face has in the image, denoted as l_f^0. Given l_f^0, d_0 and L_0 as known prior knowledge, the focal length f can be expressed as:

$$f = d_0 \frac{l_f^0}{L_0} \tag{2}$$

Once we get the measurement l_f from a frame, it should satisfy the following equation.

$$\frac{l_f}{L_0} = \frac{|A'B'|}{|AB|} = \frac{|O_C P'|}{|O_C P|} = \frac{|O_C O'|}{|O_C O|} = \frac{f}{z_P} \tag{3}$$

Note that this equation holds only if line AB is parallel to the projection plane.

Substitute f using equation (2), we can estimate the center of the human face $[\hat{x}_P, \hat{y}_P, \hat{z}_P]^T$ in the camera coordinate frame based on the coordinate $[i_P, j_P]^T$ of P':

$$\begin{aligned}
\hat{z}_P &= d_0 \frac{l_f^0}{L_0} \frac{L_0}{l_f} = d_0 \frac{l_f^0}{l_f} \\
\hat{x}_P &= \frac{\hat{z}_P (i_P - i_0)}{f} = \frac{(i_P - i_0)}{l_f^0 \cdot d_0 / L_0} \cdot d_0 \frac{l_f^0}{l_f} = \frac{L_0 (i_P - i_0)}{l_f} \\
\hat{y}_P &= \frac{\hat{z}_P (j_P - j_0)}{f} = \frac{(j_P - j_0)}{l_f^0 \cdot d_0 / L_0} \cdot d_0 \frac{l_f^0}{l_f} = \frac{L_0 (j_P - j_0)}{l_f}
\end{aligned} \tag{4}$$

The necessary measurements, i.e. distance $\hat{d}$, height $\hat{h}$ and yaw angle $\hat{\psi}$, can be calculated based on $[\hat{x}_P, \hat{y}_P, \hat{z}_P]^T$,

$$\hat{d} = \sqrt{\hat{x}_P^2 + \hat{z}_P^2},\ \hat{h} = h_0 - \hat{y}_P,\ \text{and}\ \hat{\psi} = \arcsin(\frac{\hat{x}_P}{\hat{d}}) \tag{5}$$

where h_0 is the human's height.

According to (4) and (5), the prior knowledge we need for computing distance, height and yaw angle are l_f^0, d_0 and L_0, which can be easily measured. Therefore, in our work, we do not need to calibrate the camera. The assumptions about the camera are that the focal lengths in X_c and Y_c directions are equal and the optical center of the camera is the center of the image.

Note that the estimated measurements from vision can be relatively inaccurate compared to the measurements from a 3D localization system. The inaccuracy first comes from the poor quality of the video stream. Since the camera is an analog camera, the video stream includes noise. The inaccuracy also comes from the face tracking algorithm. The face region determined by the KLT tracking may not be exactly covering the human face, so the face center position $[i_P, j_P]^T$ and face length l_f contain some errors. These issues can be compensated by well designed feedback controllers for the blimp flight.

C. Human Following Control

Blimps have dynamics that are different from quad-rotors and small airplanes. The general model of blimp has six degrees of freedom and is highly nonlinear and coupled. Based on the self-stabilized physical design of the GT-MAB, the roll angular velocity and the pitch angular velocity are negligible during the blimp flight. The blimp dynamics can be described by three simplified motion primitives:

1. Distance. The blimp can change distance along the horizontal direction that is aligned with its propellers.
$$m\ddot{d} = F_z + f_z \tag{6}$$
where d is the relative distance between blimp and human, f_z is the force generated by the two horizontal propellers and F_z is the external forces in Z_C direction.
2. Height. The blimp can ascend or descend to a desired height.
$$m\ddot{h} = F_y + f_y \tag{7}$$
where h is the height of blimp with respect to the ground, f_y is the force generated by the two vertical propellers and F_y is the external forces in Y_C direction.
3. Yaw angle. The blimp is able to spin in place so that its yaw angle can be stabilized at any desired value.
$$I\ddot{\psi} = M + \tau. \tag{8}$$
where ψ is the yaw angle, τ is the torque generated by propellers and M is the external moments exerted on blimp.

The external terms F_z, F_y and M are disturbances for the blimp and cannot be ignored. To compensate these disturbances, we introduce three feedback controllers to achieve

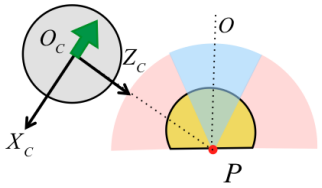


Fig. 6. An illustration of side-way motion motor on-off control from the top view. The grey circle represents the blimp and the yellow circle represents the human. The blue area is the human's front face and the red areas represent human's side faces. The green arrow represents the force generated by the side-way motion motor.

stable blimp flight based on the estimation computed by (5). The distance controller uses the estimated distance $\hat{d}$ as feedback measurement and f_z as control command. The height controller uses the estimated height $\hat{h}$ as feedback measurement and f_y as control command. The heading controller uses the estimated yaw angle $\hat{\psi}$ as feedback and τ as the control command. The goal is to control the blimp so that it keeps a constant distance d_0 away from the human at all times, in conjunction with the human moving, while keeping the human face at the center of the image, i.e. $\hat{d} = d_0$, $\hat{h} = h_0$ and $\hat{\psi} = 0$.

Because the measurements from a single camera are not accurate, the controllers need to be robust to the errors between estimated position and true position. Besides, since the blimp is required to keep the human face in sight of the camera for the entire time, the controllers need to be carefully designed such that blimp can fly fast enough to follow the motion of the human. That is to say, the settling time of each controller should be relatively short. Meanwhile, the blimp cannot move too fast as it may scare people. In other words, large overshoot of the controller should be avoided.

The controllers are designed as three PID controllers. The PID parameters are carefully tuned in MATLAB based on the system identification of GT-MAB such that all the control performance requirements mentioned above can be satisfied. The PID parameters are shown in the Table I.

TABLE I
PID CONTROLLER GAINS

Controllers	P	I	D
Distance	0.0125	0	0.0658
Height	1.3120	0.0174	1.4704
Yaw	0.3910	0	0.3840

To keep the human face in view of the blimp camera, we also use an on-off controller for the side-way motion to ensure that GT-MAB is always facing the front of human. As illustrated by figure 6, once the side face is detected, the side-way motion motor will be activated and generate a small force (green arrow) along the X_C direction. The force can regulate the blimp to fly back to the blue area, facing the front of the human. Once the blimp detects the front face of the human, the side-way motion motor will be turned off.

IV. RESULTS

We test the human following behavior with the human as the leader and the blimp as the follower through experiment. As the human moves up, down, right, left, forward and backward, the blimp is able to follow the human to a certain extent given that human is not moving too fast. In the experiment, we set the desired relative distance d_0 between the human and the blimp equal to 1.5 meters. To test the performance of our human following algorithm, we use an external real-time tracking system, OptiTrack, to measure accurate 3D position of the human and the blimp. Note that OptiTrack data is only used for analyzing the performance of our method. The data used for the human detection and the blimp control is only from the on-board camera.

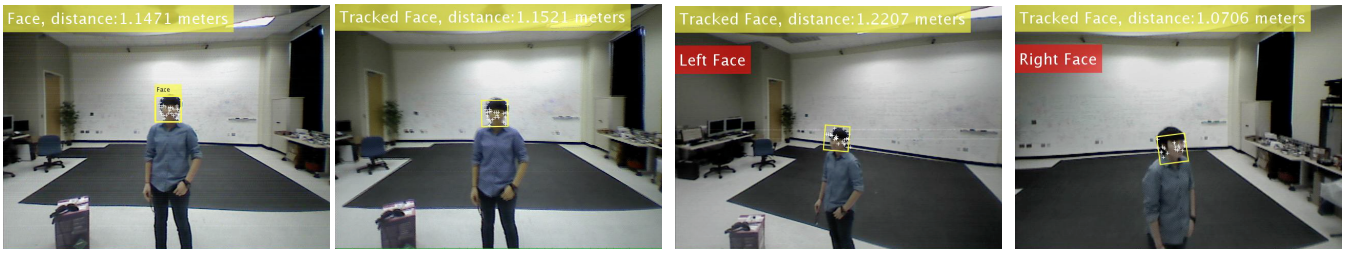
Figure 7 shows the snap shots from the blimp camera. GT-MAB can successfully track a human face and determine if it is facing the side of human face while it is flying. Figure 7(a) shows that a human face is detected by Algorithm 1. The estimated distance between the human and the blimp is also shown in the figure. Figure 7(b) shows the same face can be tracked. Figure 7(c) and 7(d) show that our algorithm can distinguish the human's left and right face.

Figure 8 shows a 3-dimensional view of the blimp and human trajectories. The blue solid line represents the trajectory of the human with the circle as the starting point and the star as the ending point. The red dashed line represents the trajectory of the blimp. The coordinate in this figure is the OptiTrack coordinate in units of meters. Figure 9 shows a top view of the blimp and human trajectories. From these two figures, we can see that the trajectories of the human and GT-MAB are similar.

Figure 10 shows the height of the blimp and human in the Z axis of OptiTrack system. The human kneels down three time to test the height control and the blimp can change its height corresponding to the human height. It appears as though the blimp is too high compared to the human. However, the OptiTrack sensors on the human are by the human's chest while the sensors on the blimp were on top of the envelope. This accounts for the difference in Z for the human and blimp. Although there is a difference in the Z position over time for the blimp and the human, the trend is the same for both, suggesting our method is accurate to some degree. Vision data from the camera is the sole measurement for carrying out our procedure. Due to this fact, our "human-following blimp" is not limited to use in a lab envirnoment but can in fact be used in other indoor environments. This is a benefit to using vision for all measurements rather than using 3D localization system such as OptiTrack.

V. CONCLUSION

Using autonomous robotic blimps as flying companions has a large potential for human-robot interaction. Human following capability is an essential prerequisite of HRI applications. This paper demonstrates the first example that our robotic blimp GT-MAB, equipped with only one monocular camera, can achieve the human following capability. This is enabled by robust human face detection and vision-based feedback control. The human following behavior is tested successfully on GT-MAB through experiments.



(a) Face detection mode (b) Face tracking mode (c) Left face detection (d) Right face detection

Fig. 7. Frames of the real-time video from GT-MAB

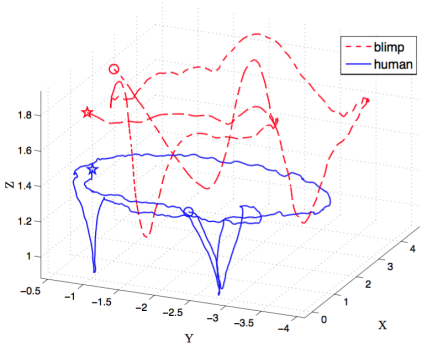


Fig. 8. 3-dimensional view of the blimp and human trajectories. The starting positions are represented by the cirles and the ending positions are represented by the stars.

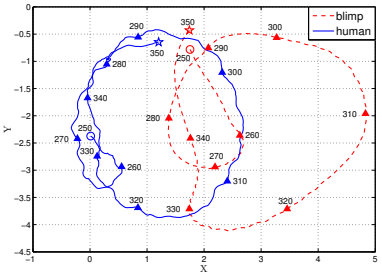


Fig. 9. Top view of the blimp (dashed) and human (solid) trajectories. The numbers in figure represent the time in the unit of seconds, showing when the human and blimp visited the points represented by the triangles.

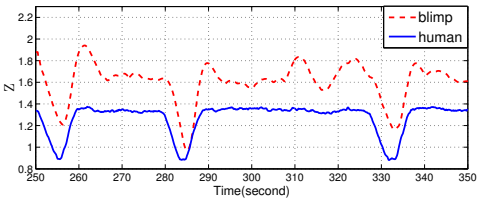


Fig. 10. Z positions of the blimp and the human

REFERENCES

[1] Q. Tao, M. King-Smith, A. Muni, V. Mishra, S. Cho, P. Varnell, and F. Zhang, "Control theory-autonomous blimp," in *IEEE CSS Video Clip Contest*, 2015.

[2] F. D. Crescenzio, G. Miranda, F. Persiani, and T. Bombardi, "A first implementation of an advanced 3d interface to control and supervise UAV (uninhabited aerial vehicles) missions," *Presence: Teleoperators and Virtual Environments*, vol. 18, no. 3, pp. 171–184, 2009.

[3] M. Draper, G. Calhoun, H. Ruff, D. Williamson, and T. Barry, "Manual versus speech input for unmanned aerial vehicle control station operations," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 47, no. 1. SAGE Publications, 2003, pp. 109–113.

[4] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[5] A. Natraj, D. S. Ly, D. Eynard, C. Demonceaux, and P. Vasseur, "Omnidirectional vision for UAV: Applications to attitude, motion and altitude estimation for day and night conditions," *Journal of Intelligent & Robotic Systems*, vol. 69, no. 1-4, pp. 459–473, 2013.

[6] Y. Kuno, M. Kawashima, K. Yamazaki, and A. Yamazaki, "Importance of vision in human-robot communication understanding speech using robot vision and demonstrating proper actions to human vision," *Intelligent Environments*, pp. 183–202, 2008.

[7] V. M. 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 *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 617–623.

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

[9] R. Li, M. Pang, C. Zhao, G. Zhou, and L. Fang, "Monocular long-term target following on UAVs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 29–37.

[10] Zero Zero Robotics, "Hover camera," http://gethover.com/, 2016.

[11] M. Quigley, M. A. Goodrich, S. Griffiths, A. Eldredge, and R. W. Beard, "Target acquisition, localization, and surveillance using a fixed-wing mini-UAV and gimbaled camera," in *Proceedings of the 2005 IEEE international conference on robotics and automation*. IEEE, 2005, pp. 2600–2605.

[12] C. F. Liew and T. Yairi, "Quadrotor or blimp? noise and appearance considerations in designing social aerial robot," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 183–184.

[13] M. Burri, L. Gasser, M. Käch, M. Krebs, S. Laube, A. Ledergerber, D. Meier, R. Michaud, L. Mosimann, L. Müri *et al.*, "Design and control of a spherical omnidirectional blimp," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1873–1879.

[14] E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer vision and image understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[15] R. C. Verma, C. Schmid, and K. Mikolajczyk, "Face detection and tracking in a video by propagating detection probabilities," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 10, pp. 1215–1228, 2003.

[16] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[17] R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee *et al.*, "Boosting the margin: A new explanation for the effectiveness of voting methods," *The annals of statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[18] S. Birchfield, "KLT: An implementation of the kanade-lucas-tomasi feature tracker," 2007.

[19] P. Corke, *Robotics, vision and control: fundamental algorithms in MATLAB*. Springer, 2011, vol. 73.

[20] S. Roelofsen, D. Gillet, and A. Martinoli, "Reciprocal collision avoidance for quadrotors using on-board visual detection," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 4810–4817.