Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

# Autonomous flying blimp interaction with human in an indoor space*

Ning-shi YAO†1, Qiu-yang TAO†1, Wei-yu LIU†2, Zhen LIU2, Ye TIAN1, Pei-yu WANG1, Timothy LI1, Fumin ZHANG†‡1

[1]*School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*

[2]*College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA*

†E-mail: nyao6@gatech.edu; qtao7@gatech.edu; wliu88@gatech.edu; fumin@gatech.edu

Received Sept. 20, 2018; Revision accepted Nov. 26, 2018; Crosschecked Jan. 8, 2019

**Abstract:** We present the Georgia Tech Miniature Autonomous Blimp (GT-MAB), which is designed to support human-robot interaction experiments in an indoor space for up to two hours. GT-MAB is safe while flying in close proximity to humans. It is able to detect the face of a human subject, follow the human, and recognize hand gestures. GT-MAB employs a deep neural network based on the single shot multibox detector to jointly detect a human user's face and hands in a real-time video stream collected by the onboard camera. A human-robot interaction procedure is designed and tested with various human users. The learning algorithms recognize two hand waving gestures. The human user does not need to wear any additional tracking device when interacting with the flying blimp. Vision-based feedback controllers are designed to control the blimp to follow the human and fly in one of two distinguishable patterns in response to each of the two hand gestures. The blimp communicates its intentions to the human user by displaying visual symbols. The collected experimental data show that the visual feedback from the blimp in reaction to the human user significantly improves the interactive experience between blimp and human. The demonstrated success of this procedure indicates that GT-MAB could serve as a flying robot that is able to collect human data safely in an indoor environment.

**Key words:** Robotic blimp; Human-robot interaction; Deep learning; Face detection; Gesture recognition
https://doi.org/10.1631/FITEE.1800587 **CLC number:** TP24

## 1 Introduction

Recent advances in robotics have enabled the rapid development of unmanned aerial vehicles (UAVs). With increasing penetration of UAVs in industry and everyday life, cooperation between humans and UAVs is quickly becoming unavoidable. It is extremely important that UAVs interact with humans safely and naturally (Duffy, 2003; Goodrich and Schultz, 2007), and to this end, the study on human-robot interaction (HRI) has enjoyed recent research interest (Draper et al., 2003; de Crescenzio et al., 2009; Duncan and Murphy, 2013; Acharya et al., 2017; Peshkova et al., 2017). Quad-rotors are one of the most popular robotic platforms for three-dimensional (3D) HRI studies (Graether and Mueller, 2012; Arroyo et al., 2014; Szafir et al., 2015; Cauchard et al., 2016; Monajjemi et al., 2016). Humans can use speech/verbal cues (Pourmehr et al., 2014), eye gaze (Monajjemi et al., 2013; Hansen et al., 2014), and hand gestures (Naseer et al., 2013; Costante et al., 2014) to command quad-rotors to accomplish certain tasks. In addition to quad-rotors,

‡ Corresponding author

* Project supported by the Office of Naval Research (Nos. N00014-14-1-0635 and N00014-16-1-2667), the National Science Foundation, U.S. (No. OCE-1559475), the Naval Research Laboratory (No. N0017317-1-G001), and the National Oceanic and Atmospheric Administration (No. NA16NOS0120028)

ORCID: Fumin ZHANG, http://orcid.org/0000-0003-0053-4224

other types of UAVs, such as fixed-wing aircrafts (He et al., 2011) and flying displays (Schneegass et al., 2014), have been developed to interact with humans.

In contrast to fast-flying UAVs, autonomous blimps are the preferred platform for HRI (Liew and Yairi, 2013) in certain applications where human comfort is a major concern. In Burri et al. (2013), a spherical robotic blimp was proposed to monitor activities in human crowds. St-Onge et al. (2017) demonstrated that a cubic-shaped blimp can fly close to human artists on stage. In another case, a robotic blimp (Srisamosorn et al., 2016) was used for monitoring elderly people inside a nursing home. These applications demonstrate the necessity of studying human-blimp interaction. However, there is a lack of dedicated design for autonomous blimps to support experiments in human interaction with flying robots in an indoor lab space.

We developed the Georgia Tech Miniature Autonomous Blimp (GT-MAB), which is designed to collect experimental data for indoor HRI (Cho et al., 2017; Yao et al., 2017; Tao et al., 2018). Being a flying robot, GT-MAB does not pose the safety threats and anxiety that a typical quad-rotor can cause to humans, and it can fly close to humans in indoor environments. In addition, GT-MAB has a relatively long flight time of up to two hours per battery charge, which supports uninterrupted HRI experiments. In this study, we introduce the hardware designs, perception algorithms, and feedback controllers on GT-MAB that identify human intentions through hand gesture recognition, communicate the robot's intentions to its human subjects, and execute the blimp behavior in reaction to the hand gestures. These features are the basic building blocks for more sophisticated experiments to collect human data and study human behaviors.

Achieving natural HRI can be more easily accomplished when the human subject does not need to wear tracking devices or use other instrumentation to interact with the robot (Duffy, 2003). With only one onboard monocular camera installed on GT-MAB, its perception algorithms can identify human intentions. We implemented a deep learning algorithm (specifically, the single-shot multibox detector (SSD) in Liu et al. (2016)), so GT-MAB could detect human faces and hands. Then we applied principal component analysis (PCA) (Wold et al., 1987) to robustly distinguish hand waving gestures in the horizontal and vertical directions. These two hand gestures trigger different reactions in the blimp. A person uses horizontal hand gestures to trigger spinning in GT-MAB and uses vertical hand gestures to trigger the blimp to fly backward (Fig. 1). We use monocular vision to measure the position of the human relative to the blimp. Vision-based feedback controllers then enable the blimp to autonomously follow a person while identifying the hand gestures. GT-MAB communicates its intentions by displaying immediate visual feedback on an onboard light-emitting diode (LED) display. The visual feedback is proven to be a key feature that improves the interactive experience.

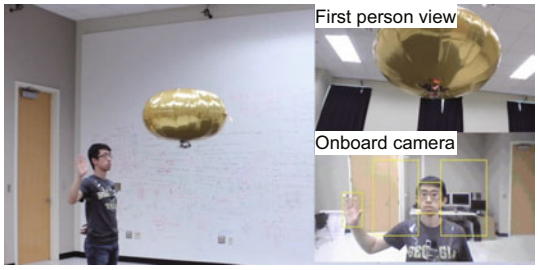


**Fig. 1 An uninstrumented human user interacts with the Georgia Tech Miniature Autonomous Blimp (GT-MAB) in close proximity and commands the GT-MAB via gestures**

We conducted HRI experiments with multiple human participants and presented a user study to evaluate the effectiveness of the proposed HRI procedure. In our experiments, GT-MAB reliably demonstrated its ability to follow humans and it consistently collected the human data while interacting with humans. In the user study, most of the participants could successfully control the robotic blimp using the two hand gestures and reported positive feedback about the interactive experience. These results clearly demonstrated the effectiveness of the basic features of GT-MAB.

## 2 Literature review and novelty

### 2.1 Data collection in the human intimate zone

Hall (1966) defined space in terms of distance to humans. The intimate (0–0.45 m), personal (0.45–1.2 m), social (1.2–3.6 m), and public (> 3.6 m) spatial zones have been widely used in both

human-human and HRI literature. However, because of their relatively high speed and powerful propellers, quad-rotors normally need to keep a relatively far distance from humans to ensure safe and comfortable interaction. In the previous HRI works (Monajjemi et al., 2013; Naseer et al., 2013; Costante et al., 2014; Nagi et al., 2014), researchers proposed similar HRI designs whereby a human user could control a single quad-rotor or a team of quad-rotors through face and hand gesture recognition. However, quad-rotors need to stay more than 2 m away from the humans to protect the users and avoid making the user feel threatened. Duncan and Murphy (2013) suggested that the minimum comfortable distance for humans interacting with small quad-rotors could not be less than 0.65 m. It is difficult for UAVs with strong propellers or the existing blimps (due to their size and functionality) to enter the human user's intimate space and collect human data without prompting anxiety on the user. GT-MAB can interact with humans within 0.4 m and collect videos of the human and the human's trajectories, which can be used to fit the social force model of Helbing and Molnár (1995) in the intimate zone. To the best of our knowledge, GT-MAB is perhaps the first aerial robotic platform that is able to collect HRI data naturally within the human intimate spatial zone.

### 2.2 Visual feedback from blimp to human

Visual feedback in the HRI procedure can significantly improve the interactive experience. Previous research has explored the implicit expressions of robot intentions by manipulating the flying motions (Sharma et al., 2013; Szafir et al., 2014; Cauchard et al., 2016). However, such implicit expressions are limited when aerial robots interact closely with humans. Explicit expressions are preferred for proximal interactions. Szafir et al. (2015) devised a ring of LED lights under the quad-rotor and designed four signals to indicate the next flight motion of the quad-rotor. A user study was conducted, where human participants were asked to predict the robot's intentions. The user study verified that the LED signals significantly improved the viewer response time and accuracy compared to a robot without the signals. However, in that work, the human participants were separated from the robot's environment by a floor-to-ceiling glass panel, so it was not an interactive environment. In our work, we discovered that immediate visual feedback is crucial for reducing human's confusion caused by the time delays between the time when a robot perceives a human command and the time when the robot initiates an action. We conducted a user study for our proposed HRI process and verified that the LED feedback significantly improves the interactive experience and efficiency.

### 2.3 Monocular vision based human localization

To localize a human, quad-rotors normally require a depth camera (Lichtenstern et al., 2012; Naseer et al., 2013). Recent works (Costante et al., 2014; Lim and Sinha, 2015; Perera et al., 2018) have also used a monocular camera on UAVs to localize humans and estimate human trajectories. Since these works used quad-rotors as the HRI platforms, one unavoidable step for monocular vision is to estimate the camera pose due to the flying mechanism of the quad-rotors, which is a challenging problem. Compared to quad-rotors, GT-MAB is self-stabilized and can fly in a horizontal plane with almost no vibration, so the pitch and roll angles of GT-MAB can be approximately viewed as staying at zero. The pose of the onboard camera is fixed. Due to this unique feature, we developed a vision-based technique to localize a human in real time from the onboard monocular camera of GT-MAB.

### 2.4 Joint face and hand detection

In previous gesture-based HRIs (Monajjemi et al., 2013; Costante et al., 2014), human face detection is necessary for distinguishing a human from other objects. Once a human face is detected, a hand detector is triggered to recognize human gestures. For each frame, two feature detectors are needed to detect different human features. The computation to run two feature detectors takes a relatively long time and is hard to implement for real-time video. To speed up video processing, feature tracking algorithms are used to track the feature detected in the previous frame (Birchfield, 1996). However, the tracking algorithms cannot consistently provide an accurate and tight bounding box around the human feature. To overcome the above-mentioned problems, we use one of the state-of-the-art object detection deep learning algorithms, SSD (Liu et al., 2016), in the context of human blimp interaction.

We also build a dataset that is efficient and adequate for training the SSD for real-time face and hand detection.

In our previous work (Yao et al., 2017), we achieved human following behavior on GT-MAB. GT-MAB was able to follow a human who was not wearing a tracking device and keep the human in sight of its onboard camera based on face detection, but GT-MAB could not react to the human. In this study, we propose a novel advancement to achieve natural interaction between a human and GT-MAB by enabling GT-MAB to recognize human intentions through hand gestures and react to human intentions through visual feedback and flying motions.

## 3 GT-MAB platform

GT-MAB consists of an envelope and a customized gondola. The envelope has a unique saucer-like shape, as shown in Fig. 1, which solves the conflict between maneuverability and stability and enhances its capability to interact with humans at a close distance. The gondola is a 3D-printed mechanical structure accommodating all onboard devices underneath the envelope. Fig. 2 depicts the structure of the gondola and indicates the main components installed on it. We use five motors for the HRI application. The vertically mounted motors are used to change the altitude, while the horizontal ones enable the blimp to fly horizontally and change the heading angle. One side-way motion motor is used to keep the blimp in the front of a human. This design enables the blimp to move in the 3D space without changing its roll and pitch angles.

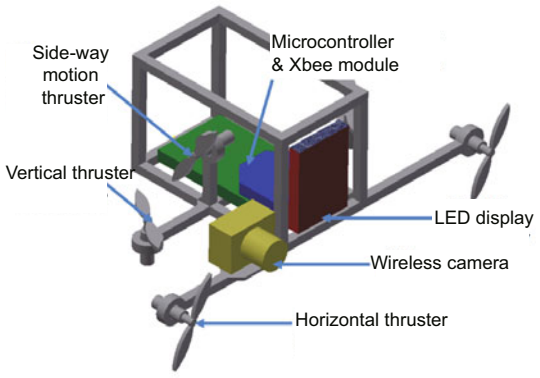


**Fig. 2 Georgia Tech Miniature Autonomous Blimp gondola with the installed electronic components**

The appealing characteristics of GT-MAB, especially the small size, impose challenges in the blimp's hardware design. The blimp has only 60 g of total load capacity, including the onboard camera, microprocessors, and wireless communication devices. One difficulty in vision-based HRI using a blimp is finding a wireless camera that is light enough. The camera we selected for GT-MAB is a 5.8-GHz analog camera, which is the best option we could find that can support low-latency wireless transmission. This compact device weighs 4.5 g and has a diagonal field of view of 115 degrees. The camera is directly attached to the gondola. However, since the camera is analog, the video produced from it includes some glitch noise, which makes image processing more difficult than with digital cameras. We also installed an $8 \times 8$ LED matrix display on the blimp to provide the visual feedback for human users. The LED display shows the recognition results, while the controller outputs achieve spinning and backward motions for the control of the blimp. Fig. 3 shows the block diagram of the hardware setup for the system. The video stream coming from the onboard camera is obtained by the receiver connected to the ground station PC. Outputs of the perception and control algorithms running on the ground PC are packed into commands and sent to GT-MAB via an Xbee wireless module.

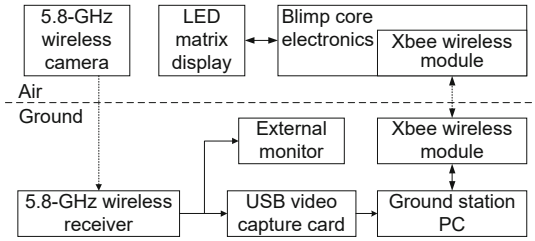


**Fig. 3 Hardware overview**

## 4 System overview

We achieve a natural and smooth HRI by enabling GT-MAB to perceive human intention. Humans are required to communicate their intentions to the blimp through predefined hand gestures so that human intentions are regulated and predictable. The human uses only one hand, starts the hand gesture near the face, and moves his/her hand horizontally or vertically. Then the blimp spins or flies backward

according to the two hand gestures.

The overall HRI design involves five steps: (1) detecting a human face and hands jointly in a real-time video stream, (2) recognizing hand gestures, (3) communicating the blimp's intentions to the human through the onboard LED display, (4) estimating the human's location relative to the blimp, and (5) controlling the blimp to follow the human and initiate movement according to hand gestures. Fig. 4 shows a block diagram of the proposed HRI design. We first run a joint face and hand detector to detect human features in each video frame. If no face or hand is detected, the onboard LED displays the negative feedback for a human and the detector goes to the next frame. If a face is detected, a human localization algorithm and a human following controller are triggered to maintain the blimp's position relative to the human. If both face and hand are detected, a gesture recognition algorithm is triggered and the LED display shows a positive feedback to the human indicating that the gesture recognition has started. Once a valid gesture is recognized, the LED display shows another positive feedback indicating that GT-MAB has received the human's command. Meanwhile, the blimp controller switches from human following control to blimp motion control, which controls the blimp to initiate spinning or backward motion. Once a motion is accomplished, the joint face/hand detector is activated again and the whole interactive process repeats. The details of each step will be introduced in the following sections.

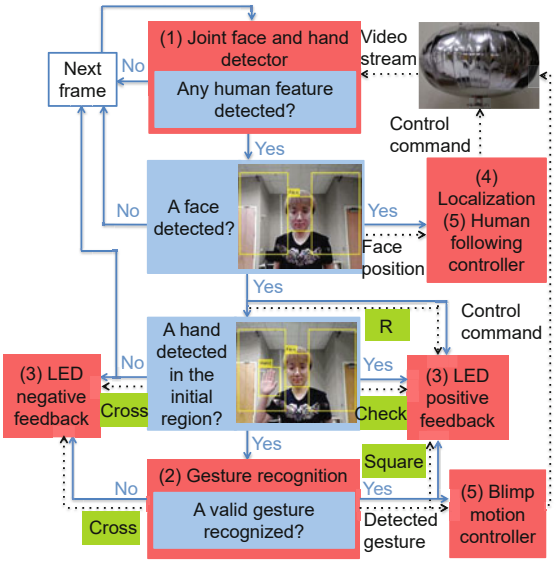


**Fig. 4 System block diagram. The pink blocks represent the five steps. The green blocks represent the visual feedback shown on the LED display. The blue blocks represent logic questions. The blue solid arrows represent the logic flow for the process and the dashed arrows represent the data flow between each block. References to color refer to the online version of this figure**

## 5 Perception algorithms

In this section, we present the first three steps of GT-MAB's perception capabilities.

### 5.1 Joint detection of face and hand

We leverage the SSD (Liu et al., 2016), which is fast and can detect multiple categories of objects at the same time to jointly detect a human face and hands in real-time videos. The idea behind SSD is simple. It reframes object detection as a single regression problem where object bounding boxes are assigned with confidence scores representing how likely a bounding box is to contain a specific object. To train the SSD, the learning algorithm discretizes a training image into $S \times S$ grid cells. Each cell has $B$ default bounding boxes of different locations and sizes. During the training process, these default boxes are compared with the ground-truth bounding boxes in training images, and a confidence score is computed for each object category. The neural network is trained to determine which default box has the highest corresponding confidence score. During detection, the trained neural network can directly generate the bounding box with the highest confidence score and determine to which category the bounded object belongs.

Particularly, to train the SSD for joint detection of a human face and hand, we create a new training set leveraging an image dataset from the Oxford Vision Group (Mittal et al., 2011). The images in this dataset have already been labeled with the human's hands. However, the human faces in this dataset are not labeled. To modify the dataset for both hand and face detection, we first assign the originally labeled hand regions as category 1. Then we use the Haar face detector (Viola and Jones, 2004) to detect a human face and label the face bounding box as category 2. We divide the modified dataset into a training set, which contains 4069 images, and a test set, which contains 821 images. The joint face and hand detector is then trained offline using the new

training set. We fine-tune the neural network using a stochastic gradient descent with 0.9 momentum, 0.0005 weight decay, and a 128 batch size. As for the learning rate, we use $4 \times 10^{-4}$ for the first $5 \times 10^4$ iterations, and then continue training for $3 \times 10^4$ iterations with a $4 \times 10^{-5}$ learning rate and another $2 \times 10^4$ iterations with a $4 \times 10^{-6}$ learning rate.

The trained joint face and hand detector is evaluated on the test set using the mean average precision (or mAP), a common metric used in feature and object detection. Specifically, for each bounding box generated by the trained detector, we discard the box if it has less than $k$ percent intersection over the union with the ground-truth bounding box. Given a specific threshold $k$, we compute the average precision (or AP) for each test image. Then we compute the mAP by taking the mean of all APs for all the test images. The test results are that with $k = 25\%$, the detector can achieve 0.862 mAP, with $k = 50\%$, the detector can achieve 0.844 mAP, and with $k = 75\%$, the detector can achieve 0.684 mAP. The performance is almost the same as that in Liu et al. (2016).

After testing the joint face and hand detector, the detector is applied to detect a human face and hand in the real-time video stream from the blimp camera. The results are shown in Fig. 5. The detected face is bounded by the yellow box with a label "Face" and the detected hand is bounded by the box labeled as "Hand." Fig. 5a shows the case where only a face is detected. Fig. 5b shows the case where both a face and a hand are detected but the hand is outside the initial gesture region, i.e., the two yellow boxes near the face bounding box. We define the initial gesture regions to filter out incorrect human gestures or random hand movements, and to ensure that the gesture recognition is more robust. Fig. 5c shows the case where both a face and a hand are detected with the hand in the initial gesture region. Only this case initializes the gesture recognition step.

Based on the bounding boxes, we define the position of the human face to be the center of the face bounding box, denoted as $\boldsymbol{P}' = [i_P, j_P]^{\mathrm{T}} \in \mathbb{R}^2$, and the face length $l_{\mathrm{f}}$ in the image frame, where $i_P$, $j_P$, and $l_{\mathrm{f}}$ are in pixels. The hand position is the center of the hand bounding box, denoted as $\boldsymbol{x} = [i, j]^{\mathrm{T}} \in \mathbb{R}^2$. We use the face position and the length of the human face to estimate the human position relative to the blimp, which will be introduced in Section 6.1.

### 5.2 Hand gesture recognition

Once the gesture recognition algorithm is initialized, the algorithm identifies two types of hand movements: horizontal linear hand movements and vertical linear hand movements.

The detection algorithm tracks the human hand from frame to frame. Once gesture recognition is triggered, the hand position is not restricted by the initial gesture region. The human hand can move out of the initial region and still be recognized. We collect the hand position data in 50 successive video frames once gesture recognition is triggered. The hand trajectory is modeled as a set of two-dimensional (2D) points $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{50}]^{\mathrm{T}} \in \mathbb{R}^{50\times 2}$ in the $O_I X_I Y_I$ image coordinates, where $\boldsymbol{x}_m = [i_m, j_m]^{\mathrm{T}} \in \mathbb{R}^2$ $(m = 1, 2, \ldots, 50)$ is a 2D vector of the hand position. If the human performs a defined gesture for the blimp, the distribution of hand trajectory data $\boldsymbol{X}$ should be close to a line. We use PCA (Wold et al., 1987) to analyze the linearity

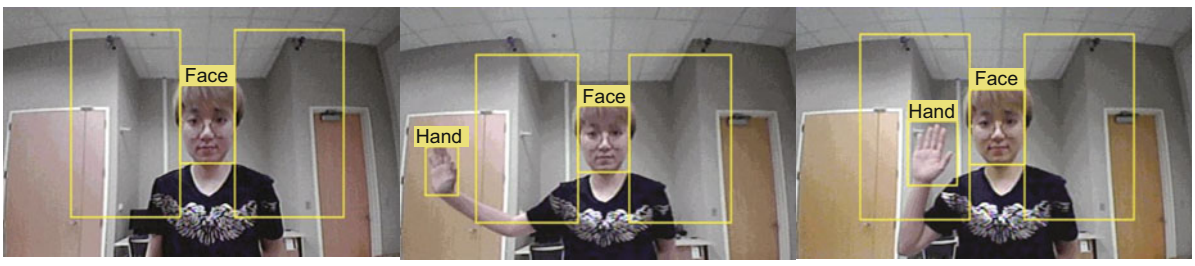


**Fig. 5 Face and hand detection: (a) only face is detected; (b) face and hand are detected with the hand outside the initial region; (c) face and hand are detected with the hand in the initial region. The images are from the onboard camera of Georgia Tech Miniature Autonomous Blimp. References to color refer to the online version of this figure**

of the data points in $\boldsymbol{X}$ and determine whether a hand trajectory is valid as defined. PCA is an orthogonal linear transformation which transforms the dataset $\boldsymbol{X}$ into a new coordinate system such that the greatest variance in the data lies on the first coordinate, and the second greatest variance lies on the second coordinate (Fig. 6). In our setup, the direction of the first coordinate from PCA is exactly the hand movement direction.

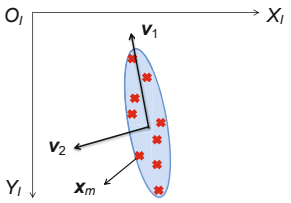


**Fig. 6 PCA illustration. The red crosses represent the data points $x_m$, $v_1$ represents the first coordinate, and $v_2$ represents the second coordinate. References to color refer to the online version of this figure**

To apply PCA, we first need to compute the mean-subtracted dataset $\boldsymbol{X'} = [\boldsymbol{x}'_1, \boldsymbol{x}'_2, ..., \boldsymbol{x}'_{50}]^{\mathrm{T}}$, since the hand positions are in pixels which are positive integers and do not have a zero mean. Each element $\boldsymbol{x}'_m$ $(m = 1, 2, \ldots, 50)$ equals $\boldsymbol{x}_m - \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is the mean of $\boldsymbol{X}$. Then the principal component can be obtained using singular value decomposition (SVD):

$$\boldsymbol{X'} = \boldsymbol{USV}^{\mathrm{T}}, \tag{1}$$

where $\boldsymbol{U}$ is a $50 \times 50$ orthonormal matrix, $\boldsymbol{V}$ is a $2 \times 2$ orthonormal matrix, and $\boldsymbol{S} = \mathrm{diag}(\lambda_1, \lambda_2)$ is a $50 \times 2$ rectangular diagonal matrix with $\lambda_1 \geq \lambda_2$. After applying SVD, we obtain the two bases of the new coordinates of PCA, $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, which are the two column vectors of matrix $\boldsymbol{V}$.

The ratio $\lambda_1/\lambda_2$ is computed to determine whether a hand trajectory is linear. A large ratio represents a high linearity. However, since humans cannot move their hands in a perfectly straight line, we need to add in some tolerance. To achieve high accuracy and robustness in gesture recognition, we run multiple trials using the blimp camera to collect both valid and invalid hand trajectories and finally select the threshold as five. Additionally, to avoid false detection of human hand gestures, we require the maximum first principal component among all the hand position data be greater than or equal to 250 (in pixels) so that the hand movement is noticeable enough that a human can recognize it. That is to say, if $\lambda_1/\lambda_2 \geq 5$ and $\max_{\boldsymbol{x}_m} \boldsymbol{x}_m^{\mathrm{T}} \boldsymbol{v}_1 \geq 250$, the hand trajectory is detected as a valid linear hand gesture.

For a valid linear hand gesture, the slope $v_{1,2}/v_{1,1}$ of the first coordinate $\boldsymbol{v}_1$ is used to determine the direction of the hand gesture, where $v_{1,1}$ and $v_{1,2}$ are the first and second elements of vector $\boldsymbol{v}_1$, respectively. If $v_{1,2}/v_{1,1} \leq 1$, the gesture is a horizontal gesture. If $v_{1,2}/v_{1,1} \geq 10$, the gesture is a vertical gesture. Otherwise, the hand gesture is invalid.

### 5.3 Visual display

However, using hand gesture recognition to activate the blimp reactive behaviors may not always work for human users. This is because there is a time delay between the time instant when the blimp detects a human and the time instant when the blimp initiates the corresponding movement. Although the time delay is only a few seconds, a human user may find the delay confusing because the person perceives no immediate reaction from the blimp. The human user may redo the hand gesture, approach the blimp to see whether the blimp is broken, or feel disappointed and walk away, even if the blimp actually recognizes the hand gesture and executes the correct action later.

Through these unsuccessful interactions, we discover that it is important for the blimp to communicate its intentions to humans. To achieve bidirectional communication between the human user and the blimp, we install an LED matrix screen on GT-MAB and it displays what the blimp is "thinking." The LED screen gives the human instantaneous feedback during the interactive process and shows the human the status of the blimp: whether it detects the user and understands his/her hand gesture. The spatially close interaction with the blimp enables the human to see the visual feedback from the LED display, and the visual feedback helps the human user take the correct action for the next step and increase the efficiency and satisfaction of the interaction.

We design four visual patterns on the LED display to represent the four intentions of the blimp (Fig. 7). The first pattern, which is the letter "R" in Fig. 7a, indicates that the user's face has been detected, and GT-MAB is ready to detect the human's

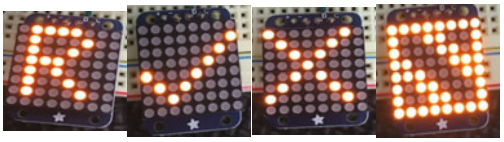


**Fig. 7 LED feedback display: (a) face is detected; (b) hand is detected; (c) a hand is not detected or a valid gesture is not recognized; (d) a valid gesture is detected, ready to fly**

hand. This is a positive feedback. When the human user sees this pattern, the human should place his/her hand near the face and start a vertical or horizontal hand movement. The second pattern, which is the "check" mark in Fig. 7b, represents that the blimp has successfully detected a human face and a hand in the initial gesture region, and it is in the process of recognizing the human's gesture. This is also a positive feedback. When the human user sees this pattern, the human should continue moving his/her hand. The third pattern, which is the "cross" mark in Fig. 7c, means that no hand has been detected in the initial gesture region or that the blimp cannot recognize a valid hand gesture. This is a negative feedback from the blimp that tells the human there was a mistake during the interaction. When seeing this pattern, the human user should place his/her hand in the initial gesture region and redo the gesture. The last pattern, shown in Fig. 7d, indicates that GT-MAB recognizes a valid hand gesture and it is going to make the corresponding motion. When seeing this pattern, the human user can see if the blimp successfully recognizes the gesture by checking whether the blimp is making the correct motion. Once the blimp completes the motion and returns to the initial position, the joint face and hand detector is triggered to detect the human face. If a face is detected, the pattern "R" is displayed again and the human can perform the next hand gesture. The whole interaction procedure repeats.

## 6 Localization and control algorithms

In this section, we present the last two steps in the HRI design for GT-MAB: vision-based human localization and blimp motion control.

### 6.1 Relative position estimation

GT-MAB localizes a human using its onboard monocular camera only. This is different from most other blimps which use an external system to localize humans, such as indoor localization or fixed external cameras (Srisamosorn et al., 2016).

We assume that the camera satisfies the pinhole camera model (Corke, 2011), which defines the relationship between a 3D point $\boldsymbol{P}=[x_P,y_P,z_P]^{\mathrm{T}}\in\mathbb{R}^3$ in the camera coordinates $O_CX_CY_CZ_C$ and a 2D point $\boldsymbol{P'}=[i_P,j_P]^{\mathrm{T}}$ in the camera image frame $O_IX_IY_I$:

$$\begin{bmatrix} i_P \\ j_P \\ 1 \end{bmatrix} = \begin{bmatrix} f_i & 0 & i_0 & 0 \\ 0 & f_j & j_0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_P \\ y_P \\ z_P \\ 1 \end{bmatrix}, \tag{2}$$

where $f_i$ and $f_j$ are the focal lengths in the $X_I$ and $Y_I$ directions respectively, and $[i_0,j_0]^{\mathrm{T}}$ is the optical center of the image. Here, we assume that $f_x$ and $f_y$ are both equal to the focal length $f$ and $[i_0,j_0]^{\mathrm{T}}$ is the center of the image.

The illustration of human position estimation is shown in Fig. 8. Because the pitch and roll angles of the 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 quadrotors because they need to change the pitch angle to fly forward or backward. GT-MAB provides a certain convenience for support of vision-based HRI algorithms because the pitch and roll angles of the onboard camera can be controlled to be zero. Line $AB$ represents the center line of the human face and we assume that it is parallel to the image plane; i.e., the plane of the human face is also perpendicular

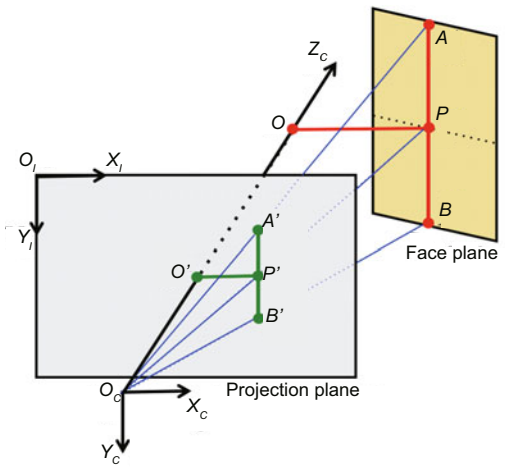


**Fig. 8 Illustration of relative distance estimation**

to the ground. Point $\boldsymbol{P} = [x_P, y_P, z_P]^{\mathrm{T}}$ is the center point of line $AB$. Points $A'$, $B'$, and $P'$ are the 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 the camera projection plane as $l_{\mathrm{f}} := |A'B'|$.

In the calibration phase, we use the detection algorithm introduced in Section 5.1 to compute a human user's face length, denoted as $L_0$ in unit of meters. The human stands away from the camera at a fixed distance $d_L$, and the position of the blimp is adjusted such that the center of the human face is at the center of the image frame. Then we run the joint face and hand detector to detect the human face and obtain the face length $l_{\mathrm{f}}^0$ in the image. Given $l_{\mathrm{f}}^0$, $d_L$, and $f$, the true human face length $L_0$ can be computed using

$$L_0 = d_L \frac{l_{\mathrm{f}}^0}{f}. \tag{3}$$

During the interaction experiments, the face length $l_{\mathrm{f}}$ from each image frame should satisfy the following equation:

$$\frac{l_{\mathrm{f}}}{L_0} = \frac{|A'B'|}{|AB|} = \frac{|O_C P'|}{|O_C P|} = \frac{|O_C O'|}{|O_C O|} = \frac{f}{z_P}. \tag{4}$$

Note that this equation holds only if line $AB$ is parallel to the projection plane. The estimated localization of the human face $[\hat{x}_P, \hat{y}_P, \hat{z}_P]^{\mathrm{T}}$ in the camera coordinate frame can be computed as

$$\begin{cases} \hat{z}_P = f\dfrac{L_0}{l_{\mathrm{f}}}, \\ \hat{x}_P = \hat{z}_P \cdot \dfrac{i_P - i_0}{f} = f\dfrac{L_0}{l_{\mathrm{f}}} \cdot \dfrac{i_P - i_0}{f} = \dfrac{L_0(i_P - i_0)}{l_{\mathrm{f}}}, \\ \hat{y}_P = \hat{z}_P \cdot \dfrac{j_P - j_0}{f} = f\dfrac{L_0}{l_{\mathrm{f}}} \cdot \dfrac{j_P - j_0}{f} = \dfrac{L_0(j_P - j_0)}{l_{\mathrm{f}}}, \end{cases} \tag{5}$$

where the true face length $L_0$ is known from Eq. (3) and the camera focal length $f$ can be obtained through standard camera calibration.

### 6.2 Blimp control

Due to the modeling (Tao et al., 2018) and the autopilot controller design (Cho et al., 2017), GT-MAB can be easily controlled to maintain its position or fly in certain patterned motions. In this subsection, we introduce three types of blimp controllers that we design for HRI application.

#### 6.2.1 Human following controller

To follow the human user and accurately track the human's hand trajectory, the goal for the human following controller is to control the blimp to maintain a fixed distance $d_0$ away from the human and to keep the human face at the center of the camera frame. The general blimp model has six degrees of freedom and is highly nonlinear and coupled. Due to the self-stabilized physical design of GT-MAB, we can use the simplified motion primitives presented in Cho et al. (2017) to design three independent PID controllers for stable human following behavior. A distance PID controller is designed to control the relative distance $\hat{d}$ to coverage to the desired value $d_0$. A height PID controller is designed to control the height difference between the human and blimp $\hat{h}$ to be 0. A heading PID controller is designed to control the difference $\hat{\psi}$ between the blimp's heading angle and the human's heading angle to be 0°. The measurements of $\hat{d}$, $\hat{h}$, and $\hat{\psi}$ can be calculated based on the estimated human position $[\hat{x}_P, \hat{y}_P, \hat{z}_P]^{\mathrm{T}}$, $\hat{d} = \sqrt{\hat{x}_P^2 + \hat{z}_P^2}$, $\hat{h} = -\hat{y}_P$, and $\hat{\psi} = \arcsin(\hat{x}_P/\hat{d})$. The PID parameters are shown in Table 1.

**Table 1 PID controller gains**

| Controller | P | I | D |
|---|---|---|---|
| Distance | 0.0125 | 0 | 0.0658 |
| Height | 1.3120 | 0.0174 | 1.4704 |
| Yaw | 0.3910 | 0 | 0.3840 |

#### 6.2.2 Blimp motion controllers

If a valid hand gesture is recognized, the blimp should not only follow the human but also make the corresponding motion controlled by the blimp motion controllers.

1. Backward motion controller

Once a vertical gesture is recognized, the backward motion controller is triggered, which also consists of three independent controllers for distance, height, and heading angle. The height and heading controllers are the same as the human following controller. The distance controller switches to an open-loop backward motion controller, which linearly increases the thrust of the two horizontal thrusters on GT-MAB until the thrust reaches its maximum limits. Under this controller, GT-MAB flies backward

(away from the human). The open-loop backward motion controller can achieve a faster and more obvious motion compared to the feedback PID controller. Once the relative distance between the human and GT-MAB reaches $(d_0 + 0.6)$ m, the backward motion is completed. The backward motion controller switches to the human following controller and then GT-MAB flies towards the human until it reaches the initial interaction distance $d_0$.

2. Spinning motion controller

Once a horizontal gesture is recognized, a spinning motion controller is activated. To achieve a spinning motion, all three PID feedback controllers for human following behavior are disabled. The spinning controller directly sets two opposite thrusts for the two horizontal thrusters so that GT-MAB can start to spin. The two opposite thrusts last 2.5 s. After 2.5 s, the horizontal thrusters stop but the spinning motion continues because of inertia. Once GT-MAB returns to its initial heading direction, the human face appearing in the video stream can be detected again and the spinning controller switches back to the human following controller.

## 7 Experiments and results

We conducted two HRI experiments on GT-MAB which validated the capabilities of GT-MAB in support of HRI. We first tested the ability of GT-MAB to follow a human and collect human data. Then we invited multiple human participants to interact with GT-MAB and collected the users' feedback to examine how participants felt about the proposed interactive procedure and how the LED visual feedback on GT-MAB affected the interactive experience.

### 7.1 Human following experiment

In this experiment, a human user served as the leader and the blimp served as the follower. As the human moved up, down, right, left, forward, and backward, the blimp followed the human to the extent that the human was not moving abruptly. We set the desired relative distance $d_0$ between the human and the blimp to 0.4 m, which regulates the blimp to follow the human in the intimate zone. We used an external real-time localization system, OptiTrack, to measure the 3D position of GT-MAB. Meanwhile, we used the localization method introduced in Section 6.1 to estimate the human trajectory online, given the position of GT-MAB from OptiTrack. To test the human following performance, we also used OptiTrack to obtain the accurate 3D position of the human, but the OptiTrack data for the human user were used only as ground truth to analyze the performance of our method. The data used for implementing human following functionality and blimp control were from the onboard camera only.

Fig. 9 shows a 3D view of the estimated human trajectory, true human trajectory, and blimp trajectory. The red solid line represents the online estimated human trajectory computed by our method. The blue dashed line represents the true trajectory of the human measured by OptiTrack. The black dotted line represents the trajectory of the blimp from OptiTrack. The coordinates in this figure are the OptiTrack coordinates in meters. Fig. 10 shows a top view of the blimp and the human trajectories. Fig. 11 shows the height of the estimated human, true human, and the blimp trajectories in the $Z$ axis of the OptiTrack system. The human kneeled down once to test the height control and the blimp can change its height corresponding to the human height. From the three figures, we can see that the estimated human trajectory and the true human trajectory matched well to a certain degree. There were some errors between the estimated trajectory and true human trajectory, even though the errors were not significant (the maximum error was 0.38 m in

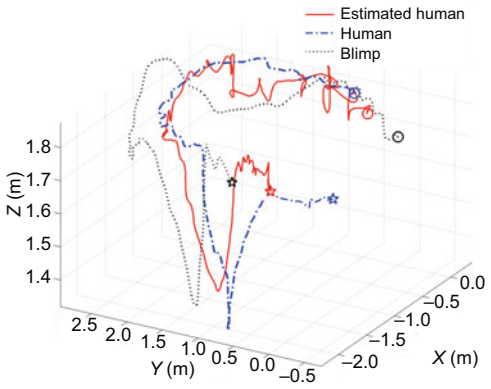


**Fig. 9 Three-dimensional view of the estimated human, true human, and blimp trajectories. The starting positions are represented by the circles, and the ending positions are represented by the stars. References to color refer to the online version of this figure**

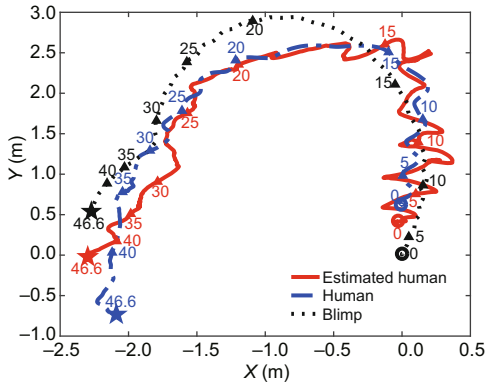


**Fig. 10 Top view of the estimated human (red solid), true human (blue dashed), and blimp (black dotted) trajectories. The numbers in the figure represent the time in the unit of seconds, showing when the trajectories visited the points represented by the triangles. References to color refer to the online version of this figure**

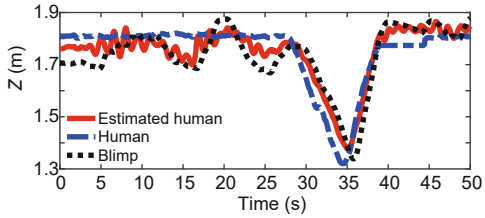


**Fig. 11 $Z$ positions of the estimated human, true human, and blimp trajectories**

the $XY$ plane and 0.11 m in the $Z$ axis). The errors were mainly due to the blimp vibrations caused by the wind from an air conditioner in our lab. Although the vibration was small and can be stabilized after a few seconds, the assumption that the camera projection plane is parallel to the human face does not hold during transitions. The errors can be reduced by conducting the experiment in a no-wind environment, or be compensated for by a low-level controller to better stabilize the vibration.

## 7.2 Evaluation of human-blimp interaction

We conducted a user study with the goal of comparing the human experience during interaction with the blimp with and without an LED display. Our goal was to verify whether the blimp's immediate visual feedback from the blimp to the human could improve the interaction.

The main hypothesis in this experiment is that human users will experience different levels of comfort with or without the LED display feedback from the blimp. This was assessed using the time duration of each interaction and a survey after the interactions.

We recruited a total of 14 participants to test the human-blimp interactive procedure designed in this study. The participants included 7 males and 7 females. The average participant age was 26.14 years old with a range of 21 to 44 years old. Six participants reported a prior familiarity with UAVs and eight participants reported a low familiarity with UAVs.

### 7.2.1 Experimental procedure

Each participant was directed to perform the procedure in a lab setting independently, i.e., without the attendance of other participants but under the guidance and supervision of the experiment assistants. We randomly separated the 14 users into two groups, groups A and B. Each group had seven participants. Participants from group A controlled the blimp without LED feedback first and with LED feedback later. Participants from group B performed the test with LED feedback first and tested without LED later. The study took approximately half an hour and consisted of three parts: pre-interaction, interaction, and post-interaction.

1. Pre-interaction

The pre-interaction began when a participant was greeted and provided consent forms with information about the study objective and his/her rights as a participant. After signing the consent forms, a participant was taken to the experiment room and guided by the experiment assistants through a few preparatory steps to learn how to interact with GTMAB: (1) An experiment assistant first played a video to the participant, demonstrating a valid horizontal hand gesture and a valid vertical hand gesture, and the corresponding blimp motions for each gesture. All participants watched the same video. (2) The assistant showed the participant pictures of the LED display patterns and informed the participant of the meaning of each pattern and what they should do after seeing each pattern. (3) The assistant demonstrated the experiment process to the participant. After the preparation, the participant was asked to practice commanding the blimp to spin and fly backward using the two valid hand gestures with and without LED display. The practice stopped when the participant felt confident that he/she could control the blimp using both hand

gestures. The practice time was less than 10 min for all participants.

2. Interaction with GT-MAB

The participants were asked to use the two gestures, horizontal and vertical hand movements. The order of the gestures could be determined by the participant. The first trial conducted by the participant was labeled as trial 1 (without LED feedback for group A and with LED feedback for group B) and the second trial conducted by the participant was labeled as trial 2. At the beginning of each trial, the participant was asked to stand at a fixed location in the experiment room and the location was unknown to the blimp. The blimp was released in front of the participant at around 1.2 m away. After the blimp was released, it automatically approached the participant and the interactive distance was set to 0.5 m, i.e., $d_0 = 0.5$ for the distance PID controller. When the blimp arrived at the desired interaction position and the human face was detected, a timer started. Meanwhile, the participant was informed by the assistant that he/she could start to perform the gesture. When the blimp recognized a valid gesture from the participant, the timer stopped. After trial 1, the participant was required to repeat the gestures to control the blimp for trial 2. The order of gestures was required to be the same as that for trial 1.

3. Post-interaction

After completing both trials, the participants were taken out of the experiment room and asked to fill out a survey form. The survey collected information including whether the participant thought the blimp took the correct action based on each gesture the participant performed, which trial brought a better interactive experience, as well as notes about the experiment in general and the interactions with GT-MAB.

### 7.2.2 Results and analysis

Experimental results showed that most of the users can interact with the autonomous blimp. There were only two gesture recognition errors among all 56 blimp control tests. The horizontal gestures from participants 7 and 9 without LED feedback were recognized as vertical gestures by the blimp. All the gestures with LED feedback were correctly recognized. This confirms that the human-blimp interaction procedure has a high success rate when used by participants who go through a short training period. We measured the amount of time that the blimp took to recognize a gesture from each trial of each participant. The time duration without LED feedback is denoted as $t_{\mathrm{NL}}$, and the time duration with LED feedback is denoted as $t_{\mathrm{L}}$. The time duration is in seconds. We also recorded videos of the participant's gestures and the blimp's corresponding motions to compare with the participant's answers collected from the survey form.

The time durations for horizontal gesture commands are shown in Fig. 12a. The blue circles represent the experimental results of participants $1, 2, 3, 7, 9, 12$, and 14 from group A, who completed the gesture without LED feedback first and with LED feedback later. The red circles represent the experimental results of participants $4, 5, 6, 8, 10, 11$, and 13 from group B, who completed the gesture with LED feedback first and without LED feedback later. Nine participants (data points below the dashed line) took less time to finish the horizontal

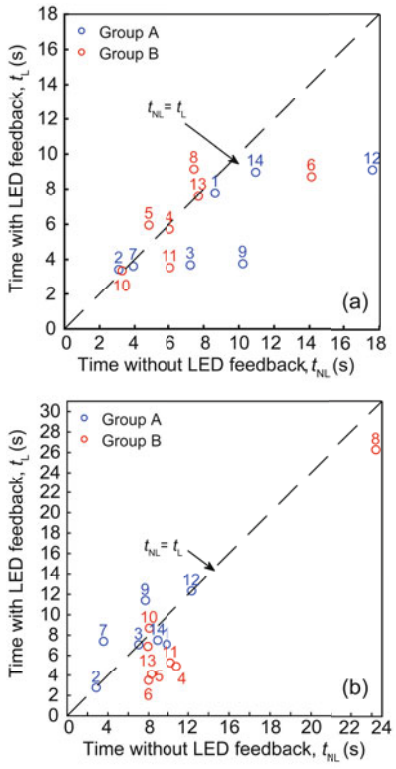


**Fig. 12 Time duration for gesture recognition: (a) horizontal gesture; (b) vertical gesture. The red circles represent the data from group A and blue circles represent the data from group B. The dashed line represents the line where $t_{\mathrm{NL}} = t_{\mathrm{L}}$. The number near each circle represents the index of each user. References to color refer to the online version of this figure**

gesture with LED feedback. Participants 10 and 13 took almost the same time to finish the horizontal gesture with and without LED feedback. Participant 2 took slightly more time (about 0.3 s) to finish the horizontal gesture with LED feedback. Participants 5 and 8 took about 1.5 s more to finish the horizontal gesture with LED feedback.

The time durations for vertical gesture commands are shown in Fig. 12b. The blue circles represent the experimental results of group A and red circles represent the results of group B. Eight participants (data points below the dashed line) took less time to finish the vertical gesture with LED feedback. Participants 2, 3, 10, and 12 took almost the same time to finish the vertical gesture with and without LED feedback. Participants 7 and 9 took more time to finish the vertical gesture with LED feedback. For both gestures, most of the participants took less time to command the blimp with LED feedback.

The average time to complete a gesture that could be successfully recognized by the blimp across the 14 participants is shown in Table 2. For horizontal gestures, the average recognition time was reduced by 2.7545 s (24.7%) with LED feedback compared to the average time without LED feedback. For vertical gestures, the average recognition time was reduced by 2.1981 s (16.1%) with LED feedback compared to the average time without LED feedback. These results confirmed that the simple visual feedback improves the interactive efficiency between the human and the blimp.

**Table 2 Average time with or without LED feedback**

| Gesture | Time (s) | |
|---|---|---|
| | Without LED | With LED |
| Horizontal | 11.1567 | 8.4022 |
| Vertical | 13.6597 | 11.4616 |

The participants were asked to choose which trial (with or without LED) provided them a better interactive experience. The preferences among all participants are shown in Fig. 13. Eleven participants out of 14 reported that the interaction with LED feedback is better. Participant 14 chose the interaction without LED feedback because participant 14 mentioned in the survey that he/she felt nervous when seeing the negative feedback from the blimp, so participant 14 preferred the interaction without LED even if the blimp might misunderstand his/her commands. Participants 2 and 10 replied that the two interactive trials provided the same HRI experience for them. From the data we collected, participant 2 took a very short time (less than 3.5 s) to complete every gesture with and without LED, so participant 2 was very effective at controlling the blimp using hand gestures. Therefore, the visual feedback was not crucial for this participant. It was a similar case for participant 10, who took almost the same time to complete each gesture. All of the other 11 participants replied in the survey that the LED visual feedback provided a better interactive experience because they knew what the blimp was "thinking."

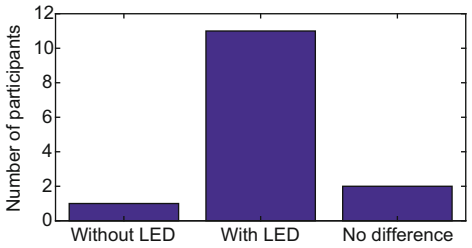


**Fig. 13 Users' preference of a better human-robot interaction experience**

## 8 Conclusions

We have presented a novel robotic platform, an autonomous robotic blimp equipped with only one monocular camera, which enables an uninstrumented human to use hand gestures to interact with the robot. The deep neural network design can effectively recognize human face and hands. The proposed learning algorithm can distinguish horizontal hand movements and vertical hand movements. The blimp reacted to humans via immediate feedback and patterned motions. We collected experimental data to show that GT-MAB has reliable human detection and human following capabilities. A user study was conducted to verify that the proposed HRI procedure can enable natural interaction between a human and a robotic blimp. We also discovered that simple visual feedback improves the interactive experience. Future work will improve the perception and learning algorithms so that more gestural commands can be interpreted by the blimp. We also acknowledged that participant groups that are more broadly representative of the potential users should be recruited to test the design of GT-MAB.

## References


Acharya U, Bevins A, Duncan BA, 2017. Investigation of human-robot comfort with a small unmanned aerial vehicle compared to a ground robot. Proc IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.2758-2765. https://doi.org/10.1109/IROS.2017.8206104

Arroyo D, Lucho C, Roncal J, et al., 2014. Daedalus: a sUAV for human-robot interaction. Proc 9th ACM/IEEE Int Conf on Human-Robot Interaction, p.116-117.

Birchfield S, 1996. KLT: an Implementation of the Kanade-Lucas-Tomasi Feature Tracker.

Burri M, Gasser L, Käch M, et al., 2013. Design and control of a spherical omnidirectional blimp. Proc IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.1873-1879. https://doi.org/10.1109/IROS.2013.6696604

Cauchard JR, Zhai KY, Spadafora M, et al., 2016. Emotion encoding in human-drone interaction. Proc 11th ACM/IEEE Int Conf on Human-Robot Interaction, p.263-270. https://doi.org/10.1109/HRI.2016.7451761

Cho S, Mishra V, Tao Q, et al., 2017. Autopilot design for a class of miniature autonomous blimps. Proc IEEE Conf on Control Technology and Applications, p.841-846. https://doi.org/10.1109/CCTA.2017.8062564

Corke P, 2011. Robotics, Vision and Control: Fundamental Algorithms in MATLAB. Springer Berlin Germany.

Costante G, Bellocchio E, Valigi P, et al., 2014. Personalizing vision-based gestural interfaces for HRI with UAVs: a transfer learning approach. Proc IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.3319-3326. https://doi.org/10.1109/IROS.2014.6943024

de Crescenzio F, Miranda G, Persiani F, et al., 2009. A first implementation of an advanced 3D interface to control and supervise UAV (uninhabited aerial vehicles) missions. *Presence*, 18(3):171-184. https://doi.org/10.1162/pres.18.3.171

Draper M, Calhoun G, Ruff H, et al., 2003. Manual versus speech input for unmanned aerial vehicle control station operations. *Proc Hum Factors Ergon Soc Ann Meet*, 47(1):109-113. https://doi.org/10.1177/154193120304700123

Duffy BR, 2003. Anthropomorphism and the social robot. *Rob Auton Syst*, 42(3-4):177-190. https://doi.org/10.1016/S0921-8890(02)00374-3

Duncan BA, Murphy RR, 2013. Comfortable approach distance with small unmanned aerial vehicles. Proc IEEE RO-MAN, p.786-792. https://doi.org/10.1109/ROMAN.2013.6628409

Goodrich MA, Schultz AC, 2007. Human-robot interaction: a survey. *Found Trends Hum-Comput Interact*, 1(3):203-275. https://doi.org/10.1561/1100000005

Graether E, Mueller F, 2012. Joggobot: a flying robot as jogging companion. Proc ACM SIGCHI Conf on Human Factors in Computing Systems, p.1063-1066. https://doi.org/10.1145/2212776.2212386

Hall ET, 1966. The Hidden Dimension. Doubleday, New York, USA.

Hansen JP, Alapetite A, MacKenzie IS, et al., 2014. The use of gaze to control drones. Proc Symp on Eye Tracking Research and Applications, p.27-34. https://doi.org/10.1145/2578153.2578156

He D, Ren HY, Hua WD, et al., 2011. Flyingbuddy: augment human mobility and perceptibility. Proc 13th Int Conf on Ubiquitous Computing, p.615-616. https://doi.org/10.1145/2030112.2030241

Helbing D, Molnár P, 1995. Social force model for pedestrian dynamics. *Phys Rev E*, 51(5):4282-4286. https://doi.org/10.1103/PhysRevE.51.4282

Lichtenstern M, Frassl M, Perun B, et al., 2012. A prototyping environment for interaction between a human and a robotic multi-agent system. Proc 7th ACM/IEEE Int Conf on Human-Robot Interaction, p.185-186. https://doi.org/10.1145/2157689.2157747

Liew CF, Yairi T, 2013. Quadrotor or blimp? Noise and appearance considerations in designing social aerial robot. Proc 8th ACM/IEEE Int Conf on Human-Robot Interaction, p.183-184. https://doi.org/10.1109/HRI.2013.6483562

Lim H, Sinha SN, 2015. Monocular localization of a moving person onboard a quadrotor MAV. Proc IEEE Int Conf on Robotics and Automation, p.2182-2189. https://doi.org/10.1109/ICRA.2015.7139487

Liu W, Anguelov D, Erhan D, et al., 2016. SSD: single shot multibox detector. Proc 14th European Conf on Computer Vision, p.21-37. https://doi.org/10.1007/978-3-319-46448-0_2

Mittal A, Zisserman A, Torr PHS, 2011. Hand detection using multiple proposals. Proc British Machine Vision Conf, p.1-11.

Monajjemi VM, Wawerla J, Vaughan R, et al., 2013. HRI in the sky: creating and commanding teams of UAVs with a vision-mediated gestural interface. Proc IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.617-623. https://doi.org/10.1109/IROS.2013.6696415

Monajjemi VM, Mohaimenianpour S, Vaughan R, 2016. UAV, come to me: end-to-end, multi-scale situated HRI with an uninstrumented human and a distant UAV. Proc IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.4410-4417. https://doi.org/10.1109/IROS.2016.7759649

Nagi J, Giusti A, di Caro GA, et al., 2014. Human control of UAVs using face pose estimates and hand gestures. Proc ACM/IEEE Int Conf on Human-Robot Interaction, p.252-253. https://doi.org/10.1145/2559636.2559833

Naseer T, Sturm J, Cremers D, 2013. FollowMe: person following and gesture recognition with a quadrocopter. Proc IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.624-630. https://doi.org/10.1109/IROS.2013.6696416

Perera AG, Law YW, Chahl J, 2018. Human pose and path estimation from aerial video using dynamic classifier selection. *Cogn Comput*, 6(10):1019-1041. https://doi.org/10.1007/s12559-018-9577-6

Peshkova E, Hitz M, Kaufmann B, 2017. Natural interaction techniques for an unmanned aerial vehicle system. *IEEE Perv Comput*, 16(1):34-42. https://doi.org/10.1109/MPRV.2017.3

Pourmehr S, Monajjemi VM, Sadat SA, et al., 2014. "You are green": a touch-to-name interaction in an integrated multi-modal multi-robot HRI system. Proc ACM/IEEE Int Conf on Human-Robot Interaction, p.266-267. https://doi.org/10.1145/2559636.2559806

Schneegass S, Alt F, Scheible J, et al., 2014. Midair displays: concept and first experiences with free-floating pervasive displays. Proc Int Symp on Pervasive Displays, Article 27. https://doi.org/10.1145/2611009.2611013

Sharma M, Hildebrandt D, Newman G, et al., 2013. Communicating affect via flight path: exploring use of the Laban effort system for designing affective locomotion paths. Proc ACM/IEEE Int Conf on Human-Robot Interaction, p.293-300. https://doi.org/10.1109/HRI.2013.6483602

Srisamosorn V, Kuwahara N, Yamashita A, et al., 2016. Design of face tracking system using fixed 360-degree cameras and flying blimp for health care evaluation. Proc 4$^{th}$ Int Conf on Serviceology.

St-Onge D, Brèches PY, Sharf I, et al., 2017. Control, localization and human interaction with an autonomous lighter-than-air performer. *Rob Auton Syst*, 88:165-186. https://doi.org/10.1016/j.robot.2016.10.013

Szafir D, Mutlu B, Fong T, 2014. Communication of intent in assistive free flyers. Proc ACM/IEEE Int Conf on Human-Robot Interaction, p.358-365. https://doi.org/10.1145/2559636.2559672

Szafir D, Mutlu B, Fong T, 2015. Communicating directionality in flying robots. Proc 10$^{th}$ Annual ACM/IEEE Int Conf on Human-Robot Interaction, p.19-26. https://doi.org/10.1145/2696454.2696475

Tao QY, Cha J, Hou MX, et al., 2018. Parameter identification of blimp dynamics through swinging motion. Proc 15$^{th}$ Int Conf on Control, Automation, Robotics and Vision. https://doi.org/10.1109/ICARCV.2018.8581376

Viola P, Jones MJ, 2004. Robust real-time face detection. *Int J Comput Vis*, 57(2):137-154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

Wold S, Esbensen K, Geladi P, 1987. Principal component analysis. *Chemom Intell Lab Syst*, 2(1-3):37-52. https://doi.org/10.1016/0169-7439(87)80084-9

Yao NS, Anaya E, Tao QY, et al., 2017. Monocular vision-based human following on miniature robotic blimp. Proc IEEE Int Conf on Robotics and Automation, p.3244-3249. https://doi.org/10.1109/ICRA.2017.7989369