2014 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS 2014)
September 14-18, 2014, Chicago, IL, USA

Human-Swarm Interaction Using Spatial Gestures

Jawad Nagi, Alessandro Giusti, Luca M. Gambardella, Gianni A. Di Caro

Abstract— **This paper presents a machine vision based approach for human operators to select individual and groups of autonomous robots from a swarm of UAVs. The angular distance between the robots and the human is estimated using measures of the detected human face, which aids to determine human and multi-UAV localization and positioning. In turn, this is exploited to effectively and naturally make the human select the spatially situated robots. Spatial gestures for selecting robots are presented by the human operator using tangible input devices (i.e., colored gloves). To select individuals and groups of robot we formulate a vocabulary of two-handed spatial pointing gestures. With the use of a Support Vector Machine (SVM) trained in a cascaded multi-binary-class configuration, the spatial gestures are effectively learned and recognized by a swarm of UAVs.**

I. Introduction

Without the use of teleoperated and hand-held interaction devices, human operators generally face difficulties in selecting and commanding individual and groups of robots from a relatively large group of spatially distributed robots (i.e., a swarm). However, due to the widespread availability of cost effective digital cameras onboard UGVs and UAVs, it is increasing the attention towards developing uninstrumented methods (i.e., methods that do not use sophisticated hardware devices from the human side) for human-swarm interaction (HSI). In previous work, we focused on learning efficient features incrementally (online) from multi-viewpoint images of multiple gestures that were acquired by a swarm of ground robots [1]. In this paper, we present a cascaded supervised machine learning approach to deal with the machine vision problem of selecting 3D spatially-situated robots from a networked swarm based on the recognition of *spatial hand gestures*. These are a natural, easy recognizable, and deviceless way to enable human operators to easily interact with external artifacts such as robots.

Inspired by natural human behavior, we propose an approach that combines face engagement and pointing gestures to interact with a swarm of robots: standing in front of a population of robots, by looking at them and pointing at them with spatial gestures, a human operator can designate individual or groups of robots of determined size. Robots cooperate to combine their independent observations of the human's face and gestures to cooperatively determine which robots were addressed (i.e., selected).

While state of the art computer vision techniques provide excellent face detection, human skeleton, and gesture recognition in ideal conditions, there are often occlusions, motion-induced blurs, and false positives, which all make the problem of detecting faces and gestures quite challenging in practice. Taking these facts into account, we consider an UAV, the A.R. Parrot flying robot as a reference model. From the one hand, this means that we are modeling in a 3D space, and from the other hand, that vision-based recognition on-board of the robot is intrinsically challenging due to the motion-induced spatial instability resulting from flying or hovering. This makes HSI interaction more challenging compared to that using ground robots. To simplify the task, we consider that the human wears a pair of colored gloves and a jacket, where the gloves are used for providing gestures and the jacket is used to detect body motion.

The main contributions of this work are: (i) a human-swarm interaction modality based on the use of spatial hand gestures given with the help of a tangible input device, (ii) the relative localization and positioning between a human and a robot swarm based on face detection and the assessment of face poses, (iii) the definition of a cascaded machine learning approach for effective spatial gesture recognition, and its use with the *distributed and cooperative classification* of hand gestures for spatial multi-robot selection.

II. Related Work

Being a relatively new area of research, human-swarm interaction aims on investigating techniques and methods suitable for interaction between humans and multi-robot systems. Existing works in HSI have adopted specific problem scenarios, which have driven the investigation of specific interfacing mechanisms and modalities [2]. Here, we direct our attention towards more general distributed sensing and recognition mechanisms that provide a swarm the distributed capability to sense audio and visual signals transmitted by human operators in the proximity.

Existing interfaces [3] for facilitating interaction between humans and multi-robot systems have typically used computer vision techniques for detecting faces, hand gestures and human body postures, where audio has been used in conjunction with vision for multi-modal interfaces. The majority of the works for selecting multiple robots using audio/video signals has been carried out by the research group of Vaughan [4], [5], [6], [7], [8]. In [4], [5] they developed a *gaze detection* approach based on face detection for selecting individuals and groups of robots in multi-robot systems. Other recent works [6], [7], [8] in this domain have adopted gaze detection as a mean of *face engagement*, for initiating interaction between humans and robots.

Recently, a multi-modal interface to select robots in a multi-robot system was presented in [7], where gaze de-

J. Nagi, A. Giusti, L. Gambardella, and G. A. Di Caro are with the Dalle Molle Institute for Artificial Intelligence (IDSIA), Lugano, Switzerland. {jawad,alessandrog,luca,gianni}@idsia.ch.

tection was used to select robots, and speech recognition was used for giving commands to robots. We consider using eye gaze at larger distances to be costly. Instead, a face can be easily detected at far greater distances. Furthermore, for localizing a human operator with multiple airborne UAVs, a visual SLAM-based approach utilizing multiple markers was adopted in [8]. To overcome the limitations of gaze detection and marker-based localization, we consider estimating the pose of the human face [9] with respect a robot's point of view, which aids human and multi-robot relative positioning.

A number of research works have promoted the use of hand gestures [10], [11] as an effective interaction tool for humans to command and control robots [4], [12], [13], [8]. Recent studies in the field of human-robot interaction (HRI) have shown that pointing gestures [14] act as a directive for robots [15] and can be used in machine vision applications [16]. Thus, in this work we adopt the use of pointing hand gestures for selecting spatially-situated individual and groups of robots, based on their spatial arrangement.

The research group of Vaughan investigated the use of gestures, however they detected human-body motion (through optical flow estimation) in specific (predefined) zones of the human body, resulting in a vocabulary of 4 motion-based waiving hand gestures [4], [17], [8]. Our proposed approach allows human operators to provide more natural and intuitive gestures, and since in our case gestures are learned and recognized based on the shape (contour) of the hand, the gesture vocabulary can be easily expanded.

In the currently existing approach for selecting groups of robots from a population [4], a human has to draw a circle (with their finger) around a desired area in front of the robots to be selected. As tracking a moving hand trajectory and determining if the face is within the hands circular motion (circumference of the drawn circle) is a complex process not suitable for use with large robot swarms, we specify the range of the group to be selected as the *spatial cone in between two pointing gestures*, which provides a more reliable and robust approach to select spatially-situated groups of robots.

III. Multi-robot Sensing and Positioning

In the following subsections III-A-III-C, we address the different aspects of the general problem we solve, starting from the definition of a basic vocabulary of static gestures for spatial selection of one or more robots from a swarm. Once the vocabulary is defined we present a method for detecting and tracking the face [18] of the human issuing the gestures. Face detection is individually performed by each robot and allows robots to identify position and visual orientation of the human operator with respect to them.

In turn, face detection is functional to determine the relative angular, radial, and altitude position of a UAV with respect to the human. Once this has been robustly assessed, robots use this information to coordinate with each other to move to positions that, if available, allow better individual views of gestures, as well as the maximization of the mutual visually sensed information at the swarm level. After this localization phase is complete, robots in the swarm start observing the human (i.e., wait for commands encoded in gestures). In particular, robots wait for spatially situated commands for selecting one or multiple robots. This is described in Section IV and its subsections.

A. Basic Gesture Vocabulary

The general objective of this work is to allow a human operator control a swarm by issuing spatial commands for selecting robots. At this aim, and for the evaluation of the techniques that will be presented in the following sections, we formulated a *basic two-handed vocabulary* of K=4 spatial gestures. This vocabulary provides spatial gestures (i.e., gestures based on angular distance) as basic commands for selecting individual robots, groups of robots, and all robots, from a robot swarm using spatial gestures. The vocabulary, as illustrated in Figure 1, satisfies the criteria of being intuitive and easy to recognize and understand for a human.

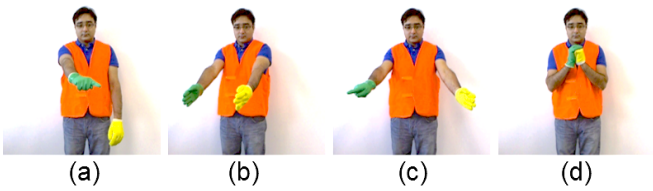

Fig. 1: The basic two-handed gesture vocabulary for spatially selecting robots. Gestures to select: (a) individual robots, (b) group of robots, (c) individuals and groups, (d) all robots.

B. Face Detection and Face Pose Estimation

We adopt the notion of *human face detection* to create a normalized and user-centric view of the human from the point of view of multiple robots. At first, the task of each UAV in our networked swarm of A.R. Parrot drones is to detect a human for interaction using its onboard camera. At this aim, we use the front-mounted cameras of the A.R. Parrot drones. Image frames acquired from the camera and flight control data are streamed using ROS onto a Linux machine via the 802.11 wireless network. Face detection is then performed using the OpenCV library implementation of the Viola-Jones face detector [19], a cascaded Haar classifier.

Inspired from the works reported in [7], [17], [4] on using face engagement for selecting and commanding robots in a multi-robot system, by setting the Haar classifier face detector parameter–the number of neighbors each candidate sub-window should retain–to be maximum, we can estimate groups of all *neighboring sub-windows* around a face. Using the *number of detected sub-windows* from a detected face (i.e., the output of the face detector, which is a measure of the quality of the detected face), the face pose of a human from a robot's point of view can be estimated [18]. We also adopt a *face pose estimation system* formulated as a non-linear regression problem which we recently developed [9]. The system uses a set of two Haar face detectors to: (i) incrementally learn symmetrical face poses (r_ϕ) of human operators through online interactions with the swarm, and (ii) robustly predict the angular distance (r_ϕ,r_d) between a

human and a robot. We exploited these properties to compute the relative distance r_d between a human and a UAV (i.e., the average area of all detected sub-windows around a detected face). Using the same technique, we also compute the face centroid (i.e., the average centroid of all detected face sub-windows) as $fc(\mathbf{x},\mathbf{y})$. We use $fc(\mathbf{x},\mathbf{y})$ as a measure for human and multi-robot positioning (see Section III-C), which helps airborne UAVs in maintaining a fixed altitude, based on the varying height of different human operators.

C. Human and Multi-robot Positioning

When robots do not know where they are located in the environment with respect to the human operator issuing commands, the correct understanding of multi-robot selection is a hard task for the robots. To allow robots to reliably learn and predict gestures based on their relative point of view from the human, it requires adjusting the spatial arrangements (positions) of the robots in the environment, prior to interaction and robot selection. In practice, it is not uncommon that some robots are not able to detect the human's face or gesture commands due to other robots obstructing their view. Therefore, to deal robustly with these potential issues, by exploiting the notion of face poses, the *face score system* developed in [9], (in Section III-B) can aid human and multi-robot positioning.

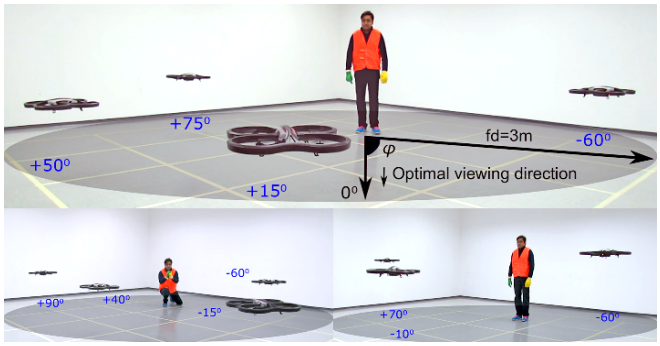

Fig. 2: Spatial arrangements of UAVs for (i) dataset acquisition, and (ii) human and multi-robot localization.

Considering a swarm of $r = \{1,2,...,N\}$ robots, we adopt the strategy presented in our previous work [20] for optimal *multi-robot (multi-sensor) positioning*, where the goal of each UAV is to move to a target position that can allow to optimize swarm's spatial distribution for the objective of obtaining robust collective recognition.

1) Angular Positioning (1st Step): With the aim of increasing the amount of the mutual information collectively gathered by the UAV swarm, the face pose r^t_ϕ predicted by each robot [9] is used to manoeuvre the *tangential position* of a robot by steering the yaw angle r^t_ϕ: as soon as the robot detects the human's face, it fixates its position in the direction facing towards the human face.

2) Radial Positioning (2nd Step): With the goal of gathering better quality observations, each UAV selects its *radial position* along a semi-circle centered around the human, as illustrated in Figure 2. This is achieved by computing r^t_ϕ and r^t_d (the distance between the human and the UAV) at every time step t, and using it as feedback for the robot's attitude controller to adjust the roll r^t_ψ and pitch r^t_θ simultaneously to let each robot place itself: (a) $\frac{180^\circ}{N}$ degrees apart from every other robot, and (b) at a distance of approximately $d = 2$m from the human. At the swarm level, this results in the maximization of the angular distance $(r_\phi, r_d)_t$ of each robot with respect to its closest neighbors.

3) Altitude Positioning (3rd Step): When interacting with robots that are close to the ground (e.g., when the UAVs are not flying), it is natural for humans to bend their body and tilt their head down. However, when the robots are airborne, their goal is to obtain high face detection estimates [18], [9]. Therefore, at each time step t, a robot checks its *elevation component* and maintains a fixed altitude r^t_α with respect to the human. This manoeuvre is performed by minimizing the Euclidean distance between the face centroid $fc(\mathbf{x},\mathbf{y})^t$ and the centroid of the acquired image, $I(\mathbf{x_C},\mathbf{y_C})^t = I(I_{w/2}, I_{h/2})^t$.

Using these *three local mobility rules*, at every time step t each UAV estimates its radial, tangential, and elevation components and steers its heading $(r^t_\theta; r^t_\psi; r^t_\phi)$ and altitude (r^t_α) in the direction provided by the resultant vector. The combined application of these rules instructs the UAV swarm to position itself along a semi-circle at regular angular intervals surrounding the human (i.e., swarm positioning for optimal sensing coverage), as shown in Figure 2 (top). The bottom left and right of Figure 2 depicts similar swarm formations using 4 and 3 UAVs respectively.

IV. Multi-robot Selection

In this section and its subsections IV-A-IV-F, we describe the second phase, which includes the sensing of the command gestures, their individual assessment by each robot, and their swarm-level final classification.

Since between issuing gestures (commands) the human operator can perform with his/her body, arms, and hands various movements which are not related to any commands, the robots first must robustly identify which gestures have a meaning for them and which do not. We deal with this issue through the definition of a *human body motion detector* based on optical flow, whose main purpose is to acquire and process only meaningful gestures defined in the vocabulary.

After every robot has understood that the human is issuing a gesture, this gesture needs to be correctly classified. A pretrained classifier is used by each robot in the swarm to produce an individual, *probabilistic opinion* regarding the gesture (i.e., the predicting the gesture). By exploiting the presence of a swarm, in which the robots in parallel acquire, process, and predict gesture images from different points of view, a *distributed consensus* algorithm is employed to fuse opinions from different views and rapidly reach a swarm-level agreement about the issued gesture.

As the gestures adopted in this work express spatially-related entities, robots in a swarm need to understand which ones among them are being selected by the human. We

address this challenge by building on the information gathered during the first phase (i.e., relative positioning between the human and the swarm), which allows to estimate the shape properties (i.e., features) of spatial gestures. Using gestures containing mutually discriminative features, we adopt a combination of machine learning, distributed information exchange, and data fusion that robustly allows to swarm to learn and recognize gestures from multiple points of view.

A. Color-based Segmentation

As we assumed that the human operator wears a pair of colored rubber gloves and a construction worker's jacket, we can conveniently perform color-based segmentation to segment both the hands and the jacket by exploiting their individual colors (yellow, green and orange respectively) in the HSV color space. After segmentation, three binary images, each corresponding to one of the three colors are obtained. Using connected component analysis we retain the largest connected component in each binary image, and remove all other smaller components (I_b^y, I_b^g, I_b^o).

The three binary images are fused together using a per-element bit-wise logical disjunction operation, that results in a single binary image I_b comprising the three segmented components (blobs) corresponding to the two hands and the jacket. Using $I_b(x,y)$, the centroid of both hands and the jacket (e.g., centroid for green glove $(C_{\mathbf{x}}^g(t), C_{\mathbf{y}}^g(t))$ at time t) is determined with respect to the x-y image plane of $I_b(x,y)$. As a final step, by tracing the contour points of the segmented jacket in $I_b(x,y)$, we calculate the lower bound y-coordinate of the jacket $J_{min}(t)$. We employ J_{min} as a measure for selecting individuals and groups of robots, (see Sections IV-F.1, IV-F.2 and IV-F.3).

B. Optical Flow Estimation of Human Motion

In order for a gesture-based interface to be fully acceptable by humans, it must allow human operators to perform gestures in the same natural way and with the same speed as they would perform gestures towards another human. Therefore, it is necessary to take into account that human operators can perform various "unnecessary" additional movements (with the hand, the arms, the body, etc.) between the moments they are issuing "significant" gesture (i.e., gesture encoded in the specified vocabulary).

The challenge in detecting *human motion* using airborne cameras comes from the fact that two sources of motion need to be taken into account. The first is caused by movements of the upper human body (i.e., hands, arms, body, and face), while the second is the due to the rapid ego-motion of airborne cameras (resulting from UAVs controlling their altitude). We address these challenges by adopting a strategy based on measures of *optical flow* to detect upper human body motion from a continuous time signal.

To obtain reliable optical flow information, we adopt a circular ring buffer to simultaneously queue (update and store) the magnitude of motion computed for the upper human body. The ring in the buffer comprises of bN elements, where N controls the amount of *damping of motion*. For every acquired image, three Euclidean distance measures, M_1, M_2, M_3, are computed and added to the the buffer. They refer to distances between the centroids of the: (i) green glove and jacket (M_1), (ii) yellow glove glove and jacket (M_2), (iii) green and yellow gloves (M_3).

Computing the difference of the Euclidean distance between consecutive frames in the buffer for all (i)-(iii) configurations, and summing the total, it determines the *magnitude of optical flow* M_t between the gloves and the jacket. At every control step t, the average optical flow magnitude, referred to as the *motion score*, is computed as $M_{score}^t = (\sum_{i=1}^{3} M_i^t)/3$. Large values of M_{score}^t means that a rapid motion is detected in the upper body region, whereas smaller values indicate there is a small fraction (i.e., motion is reduced) or no motion is detected all. In order to detect if upper body motion is present or not, we introduce a threshold parameter M_{th}. We determine $M_{th} = 1$ to be an optimal trade-off to between small fractions of motion and motions of large magnitude. Thus, when $M_{score}^t < M_{th}$, no upper body motion is present.

C. Features Extraction from Gestures

To represent spatial gestures as discriminative features for classification tasks in machine learning, we adopt an online incremental feature extraction approach by deriving a set of $N_{feat} = 30$ *geometrical shape properties* from the contour (silhouette) of the segmented yellow and green gloves (represented by I_b^y and I_b^g in Section IV-A). These features represent shapes and geometric properties that have been frequently used in literature for similar shape recognition tasks [21] and include properties such as image moments, convexity defects, roundness, aspect ratio, perimeter etc.

D. Multi-class Gesture Recognition by Single Robots

In order to learn and recognize different gestures in the vocabulary, we make use of a multi-class Support Vector Machine (SVM) classifier with non-linear Gaussian (RBF) kernel. Considering a soft-margin classification problem, we train a SVM classifier *SM* in a multi-class setting using a subset of images from the dataset (see Section III-A). After trained, the SVM supports the analytic concept of generalization and certainty, and is ready to classify a given 30-element feature vector $\bar{x}_c$ for a problem of K classes. We estimate the posterior probabilities of each gesture class using $p_i = p(y = i|\bar{\mathbf{x}})$ for $i = \{1, ..., K\}$.

For K=4 classes, a classified sample $\bar{\mathbf{x}}_c$ returns a 4-element probabilistic decision vector, $r_{\mathbf{Q}} = [q_{(i_1)}, ..., q_{(i_K)}]$, constrained by the normalization condition $\sum_{i=1}^{K} r_{\mathbf{q}_i} = 1$. As each UAV in the swarm is equipped with its own *SM* classifier, it uses this classifier *individually* predict the issued gesture by computing $r_{\hat{y}} = \arg\max_P \{q_{(i_1)}, ..., q_{(i_K)}\}$, where $r_{\hat{y}}$ is the gesture class (among the K=4 classes) with the highest probability in $r_{\mathbf{Q}}$. In this way each individual robot can effectively build an *opinion* regarding the issued gesture.

E. Distributed Consensus

In order to perform a swarm-level classification, all robots have to reach an agreement on the issued gesture. Combining individual opinions of predicted gestures with their

probabilistic scores it provides a robust strategy to boost the overall recognition performance of individual robots. To rapidly produce a collectively shared classification, we adopt the *distributed consensus* protocol developed in our previous work [20], based on the fusion of the $r_{\mathbf{Q}}$ from every robot. By exploiting the notion of distributed sensing to allow the swarm as a whole to act as a *single powerful augmented sensor*, the distributed consensus classifies a given gesture from multiple points of view with high confidence.

F. Binary-class Spatial Gesture Recognition

One of the core challenges for robots in a swarm is to determine which robot(s) the human operator is *selecting* (or trying to address) using spatial gestures, as the gesture might be visible to several robots at the same time. We address this issue by employing a soft-margin binary-class SVM classifier *SB*, that determines if the human operator is pointing towards *individual or groups of robots*.

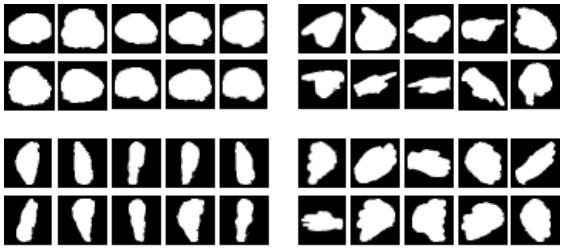

Fig. 3: Segmented images on left and right correspond to gestures (a) (b) in Figure 1. Top: Images used to train SB_I to select individual robots. Bottom: Images used to train SB_G to select groups of robots.

At this aim, we define a binary-class classification (a two-class) problem with labels $y_i \in \{-1,+1\}$, and use a subset of the images from the dataset to train two binary SVM classifiers, SB_I and SB_G for selecting respectively 'individuals' and 'groups' respectively. The segmented gestures (blobs) used to train the two classifiers are illustrated in Figure 3. The images on the top represent the two gesture classes used to train SB_I for selecting individual robots, where the top left images (i.e., one finger pointing towards a robot; *BC*=+1) correspond to one class, while the top right images (i.e., finger pointing in all other directions other than towards a robot; *BC*=-1) correspond to the other class. Similarly, the images on the bottom (that use the entire palm for pointing) belong to the two classes used to train SB_G for selecting groups.

As the binary classifiers (SB_I, SB_G) return a 2-element probabilistic decision vector $r_{\mathbf{P}} = [p_{(i_1)}, p_{(i_2)}]$ on prediction of a sample, in order to select one or more individual robots, or groups of robots from a swarm, robots make use of the results from SB_I and SB_G based on the algorithmic procedure of *distributed election* given below.

1) Selection of Individual Robots: Individual robots are spatially selected one by one (using gesture (a) in Figure 1) by using an *incremental selection* approach, as illustrated by the pseudo-code in Algorithm 1. The selection process is such that the human operator points to an individual robot and selects it, then points to another robot and selects it, and repeats the process until both the hands (gloves) of the human are below the jacket, i.e., $(C^g_{\mathbf{y}}(t), C^y_{\mathbf{y}}(t)) < J_{min}(t)$, after which the selection process is concludes.

By adopting a *distributed leader election* strategy in which the robots classify spatial gesture as pointing towards them (i.e., r_{BC}=1; see Figure 3, top left), the *individual selection score* of each robot is estimated as $r_{IS} = |p_{(i_1)} - p_{(i_2)}| \cdot (\arg\max_P \{p_{(i_1)}, p_{(i_2)}\})$. The value of r_{IS} and the robot identification number r_{id} are broadcast as an ordered pair (r_{IS}, r_{id}) to the rest of the swarm in a multi-hop fashion. To determine which robot the human is pointing at, each robot uses its list of received ordered pairs (including its own selection score), and computes $r^{win}_{ind} = \arg\max_{r_{IS}} \{(r^1_{IS}, r^1_{id}), ..., (r^N_{IS}, r^N_{id})\}$, where N represents the number of robots in the swarm and $r_{id} \in r^{win}_{ind}$ corresponds to the selected individual robot.

Algorithm 1 Incremental Selection of Individual Robots

```
1: indtotal ← 0;
2: if (λ == individual) then
3: repeatloop:
4:     while (true) do
5:         if (M_score < M_th) then //Motion detector
6:             break; //Exit while loop
7:         end if
8:     end while
9:     for i = 1 : N(r^i_IS; r^i_id) do //Selection score
10:        Compute r^i_IS from r^i_P predicted using r^i_SB_I
11:        Broadcast r^i_IS, r^i_id to robots with r^i_BC = 1
12:    end for
13:    //Highest score selected
14:    r^win_ind = arg max_rIS {(r^1_IS, r^1_id), ..., (r^N_IS, r^N_id)};
15:    indtotal++; //No. of selected robots
16:    Pause for X seconds; //Create small delay
17:    if (C^g_y > J_min) then //Check height
18:        goto repeatloop;
19:    end if
20: end if
```

2) Group Selection: For selecting groups (or teams) of spatially-situated robots from a swarm, we adopt a *simultaneous selection* approach as illustrated by the pseudo-code in Algorithm 2. At this aim, a human operator provides gesture (b) in Figure 1, which defines a "spatial cone" (or range) with the use of two hands. In this context, we consider that robots spatially located within the proximity of both hands are part of one entire group. As each UAV in the swarm uses its individual classifier SM_G to predict if the left and right hand (palm) are pointing directly towards it (see Figure 3, bottom left), the two UAVs which individually have the best view of the right and left hand pointing towards them respectively, are marked as boundaries for group selection. In simpler words, both of the hands of gesture (b) in Figure 1 are used to define the boundaries of a confined spatial area, where all robots within these boundaries constitute a group.

In order to be robust, the group selection process must ensure that when selecting subgroups of spatially-situated robots, only robots that are within physical proximity of

each other, as well as within the spatial area defined by the boundaries of both the hands, get selected as a group. At this aim, only the robots who have classified the gesture as pointing towards them (i.e., r_{BC}=1) go through a two-stage selection process. First, all robots who classify both hands as pointing towards them (see Figure 3, bottom left) according to the binary classifier r_{SB_G}, compute two individual *boundary selection scores* r_{BSy} and r_{BSg} (for the yellow and green gloves respectively) using Algorithm 1 two times (once for each hand) This means that, using an individual robot selection approach, r^{win}_{BSy} and r^{win}_{BSg} are obtained separately for both hands (gloves).

Algorithm 2 Simultaneous Selection of Groups of Robots

```
intialselection ← 0; grptotal ← 0; G_avg ← 0;
if (λ == group) && (M_score < M_th) then
    Obtain (r^win_BSy; r^id) and (r^win_BSg; r_id) using Algorithm 1
    if r_φ(r^win_BSy, r^id) < r_φ(r^win_BSg, r^id) then
        φ_min = r_φ(r^win_BSy) //Angle adjustment
        φ_max = r_φ(r^win_BSg)
    else
        φ_min = r_φ(r^win_BSg)
        φ_max = r_φ(r^win_BSy)
    end if
    for i = 1 : N do //Angular distances
        Compute r^i_φ for each robot in swarm
        Broadcast (φ_min; φ_max) to swarm as 2-tuple
    end for
    for i = 1 : N do //Group selection
        if ((r^i_φ ≥ φ_min)&&(r^i_φ ≤ φ_max)) then
            Include robot r^i_id in r̄^win_grp
            grptotal++;
        end if
    end for
end if
```

In order to estimate the spatial area (range) between the left and right robots (i.e., the group boundaries), in the second stage we identify the robots (r_{id}) that have been selected as r^{win}_{BSy} and r^{win}_{BSg}, and predict r_ϕ for both these robots (using the face pose estimation system developed in [9]; see Section III-B) with respect to the human on a horizontal [0,180°] plane. We represent the smaller angle as ϕ_{min} and the larger angle as ϕ_{max}. Finally, all robots compute their individual r_ϕ, and the robots that lie within the spatial area defined by the closed interval $[\phi_{min}, \phi_{max}]$ are selected as one entire group, denoted by $\bar{r}^{win}_{grp}$.

3) Individual and Group Selection: To select individuals and groups of robots together (see gesture (d) in Figure 1), we adopt an *incremental and simultaneous selection* approach by using both Algorithm 1 and 2. To make things simpler we consider that, one hand (e.g., yellow glove) only gives gestures for selecting groups, while the other hand (green glove) only gives gestures for selecting individuals. After a group of robots and an individual robot have been selected, selection terminates if $(C^g_{\mathbf{y}}(t), C^y_{\mathbf{y}}(t)) < J_{min}(t)$ (i.e., both hands are lower than the jacket's lower-bound y-coordinate $J_{min}(t)$). However, if one hand (yellow glove) is higher than $J_{min}(t)$, then more individual robots are selected, while if the other hand (green glove) remains higher then more robot groups are selected.

In situations when the distributed consensus identifies that *all robots* are selected, as indicated by gesture (d) in Figure 1, all robots provide a *feedback response* to the human that they have been selected by flashing their onboard LED lights. This conveys basic swarm-level information to the human operator using minimal communication complexity with immediate impact. Alternatively, locally coordinated movements can also be used to provide feedback to the human operator. In the case of individual or group selection, only selected robots provide feedback to the human operator and subsequent commands (e.g., to perform a task) are only executed by the subset of selected (engaged) robots.

V. Experimental Results

To demonstrate and quantify the capabilities of the developed system [1], we performed experiments to investigate performance, robustness, and efficiency of the solutions proposed in Sections III and IV. At this aim, we first built a dataset of images using a small swarm of 4 airborne A.R. Parrot drones equipped with front-mounted cameras capturing images at a resolution of 1280 × 720 pixels. Using the 4 drones we acquired a relatively large amount of images of our gesture vocabulary from multiple points of view. During dataset acquisition, all acquired images were labeled (tagged) with their known ground truth information, which was used to train the gesture classifiers used by the robots.

To acquire the dataset, the UAVs are positioned around the human using the multi-robot formation illustrated in the top of Figure 2. Using this configuration, each robot acquired and stored approximately 800 unprocessed images while the human operator for a short time presented gestures directed towards the robot with the most frontal (optimal) view of the human face. Given that our vocabulary consists of K=4 gestures, in total the swarm roughly acquired 4 × 800 × 4 images. This process was repeated 5 times, once for a different distance $D = \{1,2,3,4,5\}$m between the UAVs and the human, which resulted in a dataset of approximately 64,000 images acquired by the UAV swarm from a total of 4 × 5 = 20 different viewpoints.

A. Sensitivity in Human Motion Detection

To ensure that human and multi-robot interaction is natural as possible, we study the effect of using different motion damping values bN on M^t_{score}, the results are reported in Figure 4. The results show that, if bN is too small, M^t_{score} changes very rapidly and is unstable (i.e., the motion score fluctuates) and too sensitive to be used to detect human body motion reliably over time. Instead, if bN is too large, the computed optical flow is slower, meaning that motion will not be detected on the spot it occurs, but after some delay. In simpler words, small values (e.g., bN=5) indicate a fast

[1] A demonstration video of the entire system can be viewed here: `http://goo.gl/oT60Ln`. In the video, selected robots *lift-off*, *move* and *land*, similar to the use of "force" in Starwars movies.

decay in M_{score} (spikes), whereas larger values (e.g., bN=20) provide a slower decay rate (steps).

If motion is too rapid or too slow, the way robots perceive the gesture would cause color-based segmentation errors due to blur in the images making the system potentially unreliable. Therefore, choosing a good estimate of the buffer size is critical to support the reliability of the motion detection system. At this aim, we determine using bN with values in the range of $\{9,..,12\}$ provides more a smoother distribution for detecting upper body motion.

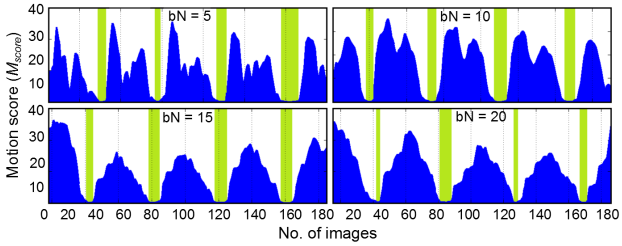

Fig. 4: Average magnitude of motion M^t_{score} for different values of the motion damping parameter bN.

B. Accuracy of Spatial Robot Selections

In order to validate our proposed solutions, we performed several experiments using different spatial configurations of individuals and groups of robots. In every configuration, a human operator attempts to select individual robots (see Figure 5) or groups of robots (see Figure 6). This is emulated by selecting subsets of images from the acquired dataset based on ground truth information (spatial arrangements of robots) and implementing Algorithm 1 and 2. The results reporting the performance and accuracy of robot selection are presented as grayscale colormaps. All reported experiments are averaged over 1000 trials, using images from similar spatial configurations of the robots on each trial.

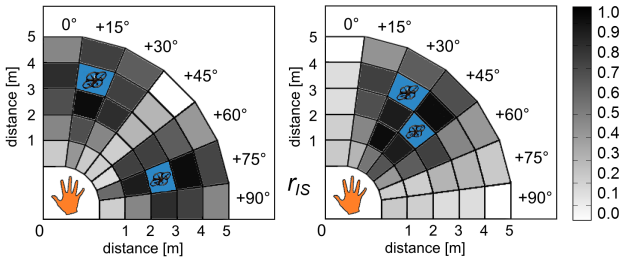

Fig. 5: Gesture classification accuracy for individual robots in different spatial configurations.

1) Effect of Selection Score on Individuals: First, we study the effect of the individual selection score $r_{IS} \in [0,1]$ on surrounding, non-selected robots. In particular, we consider two different cases (configurations) in which two individual robots are selected. In the first case, the two robots are very close to each other (see Figure 5, right) while in the second case they both robots are far apart (see Figure 5, left) from each other. All surrounding robots are uniformly spread around in the environment.

The gray colormap illustrates the individual selection scores r_{IS} for all robots surrounding the selected two robots, where positions (cells) with dark colors represent surrounding robots with large values of r_{IS} (as they are very near to the selected robots) and surrounding robots with light color cells represent that they are far from the selected robots. These results show that when robots are within close proximity of each other the success rate of selecting individual robots decreases, as expected. This can be avoided by maximizing the angular distance between each robot, as discussed in the positioning rules in Section III-C).

2) Sensitivity of Boundary Selection Scores: Secondly, we investigate the sensitivity of the boundary selection scores r_{BSy} and r_{BSg} on surrounding, non-selected robots. At this aim, we consider a swarm of N = 14 robots with spatial configurations of robots surrounding the human, as depicted in Figure 6. A group of 8 robots (located in the cells with the blue background) are selected from the swarm using both hands, while the non-selected robots are placed uniformly in each one of the remaining cells. The gray colormap presents the boundary scores for all deployed robots. Dark color cells represent surrounding robots with similar boundary scores to that of the group being selected, while surrounding robots in white cells indicate that they are far from the selected group.

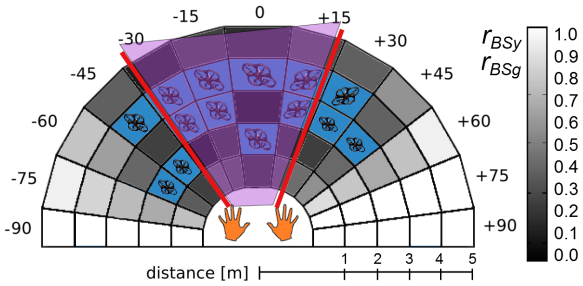

Fig. 6: Gesture classification accuracy for selecting a group (team) of robots from a swarm.

The success rate in selecting a subgroup of spatially-situated robots strongly depends upon: (a) the angular distance between the robots to be selected, and (b) the distance to other surrounding neighbors. In situations where a group of robots that has to be selected is in close proximity to other groups or individuals, there are high chances that selection may be incorrect. This is because, in close proximity, r_{BSy} and r_{BSg} of nearby robots is similar. To avoid such situations, the angular position between each robot must be 15° or more apart from each other.

C. Effect of Swarm Size on Selection and Recognition

Lastly, we investigate the effect of the size of the swarm N on the gesture recognition accuracy, as well as on multi-robot selection accuracy, as shown on the left and right of Figure 7 respectively. The general impact of an increased swarm size in relation to the recognition accuracy of the multi-class SVM (SM) and the two binary SVMs (SB_I (individuals) and SB_G (groups)), is illustrated on the left of Figure 7. Increasing the number of robots in the swarm has a positive effect on the overall gesture recognition performance. Using a swarm size of $N >= 5$ robots, gesture classification accuracy obtained

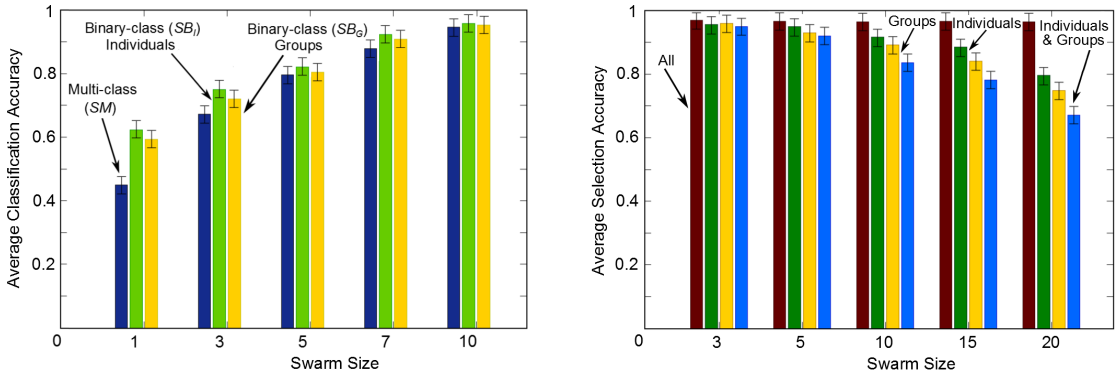

Fig. 7: Left: Effect of gesture classification performance on the size of the swarm. Right: Impact of individual and group selection accuracy on the size of the swarm.

from both the multi-class and binary SVMs are greater than 80%. This is expected, as larger swarm always result in higher recognition accuracy.

The results shown on the right of Figure 7 report the effect of the swarm size N on the multi-robot selection accuracy. It is observed that increasing the number of robots has a negative effect on the selection accuracy (i.e., the selection accuracy decreases as the swarm size increases). As a result, relatively large and densely populated swarms (e.g., N=20) are not robust towards multi-robot selection. Also, the average selection accuracy of individual robots is marginally better than that of groups and individuals and groups, due to the fact that individual robots have a wider spatial workspace, (i.e., selection of individuals is robust to a wider set of mutual poses). In simpler words, the spatial configurations (arrangements) of individual robots have a reduced effect in relation to the swarm size.

VI. Conclusions

We presented an integrated vision-based approach for the problem of selecting individual and groups of robots from a robot swarm using spatial gestures given by a human operator. The experimental results, obtained in emulation using real data acquired from a group of A.R. Parrot flying drones indicate that the proposed approach for multi-robot selection and cooperative spatial gesture recognition is robust and scales well with swarm sizes of up to 20 robots. Future work will focus on developing a self-contained grammar-based vocabulary of gestures that can spatially address, intuitively provide commands, give directions, and naturally represent quantities, for supporting fully bidirectional interaction between humans and robot swarms.

Acknowledgments

This research was supported by the Swiss National Science Foundation (SNSF) through the National Centre of Competence in Research (NCCR) Robotics (`www.nccr-robotics.ch`).

References

[1] J. Nagi, A. Giusti, F. Nagi, L. M. Gambardella, and G. A. D. Caro, "Online feature extraction for the incremental learning of gestures in human-swarm interaction," in *Proc. of IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014.

[2] A. Rule and J. Forlizzi, "Designing interfaces for multi-user, multi-robot systems," in *Proc. of the Intl. Conf. on HRI*, 2012, pp. 97–104.

[3] G. Jones, N. Berthouze, R. Bielski, and S. Julier, "Towards a situated, multimodal interface for multiple uav control," in *Proc. of IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2010, pp. 1739–1744.

[4] A. Couture-Beil, R. Vaughan, and G. Mori, "Selecting and commanding individual robots in a multi-robot system," in *Proc. of the Canadian Conf. on Computer and Robot Vision (CRV)*, 2010, pp. 159–166.

[5] B. Milligan, G. Mori, and R. Vaughan, "Selecting and commanding groups in a multi-robot vision based system," in *Proc. of ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI)*, 2011, pp. 415–415.

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

[7] S. Pourmehr, V. M. Monajjemi, R. T. Vaughan, and G. Mori, "You two! take off!: Creating, modifying and commanding groups of robots using face engagement and indirect speech in voice commands," in *Proc. of IEEE/RSJ IROS*, 2013, pp. 137–142.

[8] V. Monajjemi, J. Wawerla, R. Vaughan, and G. Mori, "HRI in the sky: Creating and commanding teams of UAVs with a vision-mediated gestural interface," in *Proc. of IEEE/RSJ IROS*, 2013, pp. 617–623.

[9] J. Nagi, G. A. D. Caro, A. Giusti, and L. M. Gambardella, "Learning symmetric face pose models online using locally weighted projectron regression," in *Proc. of IEEE Intl. Conf. on Image Proc. (ICIP)*, 2014.

[10] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. on Systems, Man, and Cyb.–Part C*, vol. 43, no. 3, pp. 311–324, 2007.

[11] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan, "Vision-based hand-gesture applications," *Comm. of ACM*, vol. 54, no. 2, pp. 60–71, 2011.

[12] K. Konda, A. Königs, H. Schulz, and D. Schulz, "Real time interaction with mobile robots using hand gestures," in *Proc. of ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI)*, 2012, pp. 177–178.

[13] T. Naseer, J. Sturm, and D. Cremers, "FollowMe: Person following and gesture recognition with a quadrocopter," in *Proc. of IEEE/RSJ IROS*, 2013, pp. 624–630.

[14] A. Sauppé and B. Mutlu, "Robot deictics: How gesture and context shape referential communication," in *Proc. of the ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI)*, 2014, pp. 342–349.

[15] S. Abidi, M. Williams, and B. Johnston, "Human pointing as a robot directive," in *Proc. of ACM/IEEE Intl. Conf. on HRI*, 2013, pp. 67–68.

[16] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human-robot interaction," *Image Vision Computing*, vol. 25, no. 12, pp. 1875–1884, 2007.

[17] A. Couture-Beil, R. Vaughan, and G. Mori, "Selecting and commanding individual robots in a vision-based multi-robot system," in *Proc. of ACM/IEEE Intl. Conf. on HRI*, 2010, pp. 355–356.

[18] J. Nagi, A. Giusti, G. A. D. Caro, and L. M. Gambardella, "Human control of uavs using face pose estimates and hand gestures," in *Proc. of ACM/IEEE Intl. Conf. on HRI*, 2014, pp. 252–253.

[19] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[20] A. Giusti, J. Nagi, L. M. Gambardella, and G. A. D. Caro, "Cooperative sensing and recognition by a swarm of mobile robots," in *Proc. of the 25th IEEE/RSJ IROS*, 2012, pp. 551–558.

[21] I. Valavanis and D. Kosmopoulos, "Multiclass defect detection and classification in weld radiographic images using geometric and texture features," *Expert Syst. with App.*, vol. 37, no. 12, pp. 7606–7614, 2010.