Survey on Natural Interaction Techniques for an Unmanned Aerial Vehicle System

Ekaterina Peshkova[1], Martin Hitz[1], Bonifaz Kaufmann[2]

[1]Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
[2]Internet of Things 40 Systems GmbH, Klagenfurt, Austria
{Ekaterina.Peshkova,Martin.Hitz}@aau.at,Bonifaz.Kaufmann@gmail.com

Abstract. The paper provides an overview of existing interaction techniques for controlling Unmanned Aerial Vehicle (UAV) systems. This work focuses on user interfaces with non-traditional input modalities such as gestures, speech, and gaze direction. Although we analyze interaction with UAV systems, most of the findings can be applied to Human-Robot Interaction in general. We report on interaction techniques employed to control single as well as multiple UAV systems, define *intuitiveness* of input vocabularies in the considered context, and introduce a new classification scheme based on the mental models underlying the interaction vocabulary.

Keywords: User-centered design, Interaction techniques.

1 Introduction

Controlling an Unmanned Aerial Vehicle (UAV) system is challenging and in many cases claims the operator's constant attention and guidance. A high operator's workload is one of the major factors causing air accidents [1]. The deliberate development of user interfaces for UAVs which significantly reduces the high workload for operators, especially in a multi-UAV case, is crucially important.

A potential way to overcome the problem of an unnecessarily high workload is the utilization of natural and intuitive interaction techniques. Apparently, natural modalities such as speech and gestures have the potential to bring naturalness to Human-Robot Interaction, while a proper vocabulary might lead to a greater intuitiveness of an interface. This leads to several questions: How do we develop an intuitive input vocabulary? Why may an input vocabulary turn out to be counterintuitive, even though each individual entry seems to make sense? In order to address these questions, we leverage the concept of mental models and apply it to user interfaces for Human-UAV Interaction by clustering supportive examples into three categories: *Imitative*, *Instrumented*, and *Intelligent*.

The remainder of this paper is organized as follows. Section 2 overviews existing natural interaction techniques, defines the notion of intuitiveness based on the concept of mental models, introduces a new classification scheme for input vocabularies, and clusters these vocabularies in accordance with the introduced classification scheme. Subsections 2.1-2.3 provide examples of corresponding classes of mental models, analyze previous work in terms of intuitiveness, and outline questions that need to be addressed. Subsection 2.4 discusses aspects related to the development of interaction techniques. Section 3 concludes and discusses possible directions for future research.

2 Interaction Techniques

UAVs that have originally been used in military missions are starting to be used in various civil applications such as surveillance, search and rescue, and transportation. In the scope of this paper, we deliberately focus on civil applications and novice users. The existence of a wide range of applications where the use of UAVs is beneficial,

IEEE Pervasive Computing, volume 6, issue 1, January-March 2017
URL: http://ieeexplore.ieee.org/document/7807193/
DOI: 10.1109/MPRV.2017.3

together with various interesting scientific challenges associated with them (e.g., flight stabilization, navigation, and coordination) constitute the key reasons for the booming scientific interest in UAVs. In addition, the recent commercial availability of low-cost UAVs is making UAV systems affordable to a larger range of researchers and practitioners.

Depending on the application of a UAV system, existing interfaces vary significantly. From entire Control Ground Stations (CGSs), in which often more than one operator is involved, and standard 'Windows, Icons, Menus, Pointer' (WIMP) interfaces that require extensive training for an operator to become professional, to simpler remote controllers and touchscreen-based interfaces that also require preliminary instructions. The emergence of new technologies creates the possibility of bringing current interaction techniques to the next level. Recently, much work has been carried out on the development of immersive flight control using advanced video feedback (e.g., Google and Epson Moverrio glasses, Oculus Rift) and natural flight control with gestures using various sensing devices to capture data for gesture recognition (e.g., Kinect, Leap controller, Myo armband).

Along with the applications listed above, where experienced human operators control the flight, nowadays, we observe a growing trend towards the development of systems where novice users interact with UAVs [2]. This defines the need for a user interface that provides an easy and fast way to interact with a system without extensive training. One way to reduce the time for preliminary preparation and ease operators' work is to develop a 'natural' user interface. It is a widely held view that the use of natural cues peculiar to Human-Human interaction could contribute to the development of a more natural Human-UAV Interaction. Over the past decade, researchers have shown increased interest in the development of natural interaction techniques using non-traditional input modalities such as speech, gestures, and gaze direction. In particular, gestures have received special attention and various gesture-based input vocabularies were suggested, including hand [3,4], head [5,6], and upper body movements [7]. Apart from gestures, research has started exploring gaze direction [8], face pose [9,10,11,12], and even brain activity [13] as potential input modalities.

Usually, researchers employ elicitation techniques such as Wizard of Oz sessions and interviews to let users define the input vocabulary. In this way, researchers aim at approaching intuitive interaction with a system. Up to now, a few authors have begun to explore users' natural behavior (speech and gestures) for steering a group of UAVs [14] and a single UAV [15,16] using Wizard of Oz sessions. Burke and Lasenby [17] conducted interview sessions to gather users' gesture suggestions for steering a UAV. In the remaining works on natural interaction

Classification schemes for gesture-based interaction

In gesture studies, many existing classification schemes of gestures originated from Efron's work [24] distinguishing gestures whose meaning is dependent (*deictic, physiographic, emblematic*) or independent (*batons, ideographic*) of speech. Among them, McNeil's classification [25] is one of the most frequently referred to. McNeil outlined gestures that visualize what is being said (*iconic* and *metaphoric*) and support the flow of speech (*beat* and *deictic*). Ekman et al. [26] outlined also *manipulators* that refer to unintentional movements (e.g., scratching), *regulators* that serve to maintain contact (e.g., head nods), and *emotional expressions*. Kendon [27] classified gestures based on their formality and speech-dependency, starting from the least formal that hardly can be correctly interpreted without spoken language (*gesticulation*) to those that can be understood independently from speech (*language-like*, *pantomimes*, and *emblems*) and ending with the most formal (*sign languages*).

The early gesture studies considered gestures that are used to enhance speech. In the field of Human-Computer Interaction (HCI), gestures are often used without speech and serve for different from narrative purposes (e.g., manipulate objects on a touchscreen, navigate a vehicle). Thus, the existing classifications could not be directly applied to gestures used in modern interactive systems.

During the last decade, we have observed the emergence of new classifications for specific domains of HCI such as surface computing [28] and 3D motion gestures for smartphones [29]. In this work, we introduce a new classification of mental models associated with the input vocabulary that is the first attempt to classify natural input vocabularies in the field of Human-UAV Interaction. Different from previous works that classify individual gestures, we use a concept of mental models to classify the entire input vocabulary. Input vocabularies for the considered interaction can include any free-space gestures as well as gaze direction, facial expressions, and speech.

techniques with UAVs considered in this paper, the researchers first suggested an input vocabulary and only then let users test it.

Once users' suggestions are collected, the next step is to select the 'best' of them for the final vocabulary. Typically, researchers either select the most frequently observed users' suggestions [18] or they choose the suggestions with the highest guessability score [19]. However, none of these two selection approaches guarantees coherence of the obtained vocabulary. In a general sense, a vocabulary is *coherent* if there is a logical relationship between its components. A possible way to evaluate the coherence of an input vocabulary is to reveal one command and ask a user to guess the remaining ones. For instance, a person is told that moving the right hand up requests a UAV to fly up. Then, keeping this hint in mind, the person should be able to guess most of the remaining basic motion commands (Figure 1, *Hand*). In this example, the person simply imitates UAV movements with corresponding hand movements. Thus, all the gestures are related based on a single metaphor that implies imitation of UAV movements. For our approach, we consider a gesture set to be *coherent* if all its gestures adhere to *one and the same* metaphor. This metaphor evokes a certain mental model that, in turn, defines a certain behavior or, in the considered example, certain gestures.

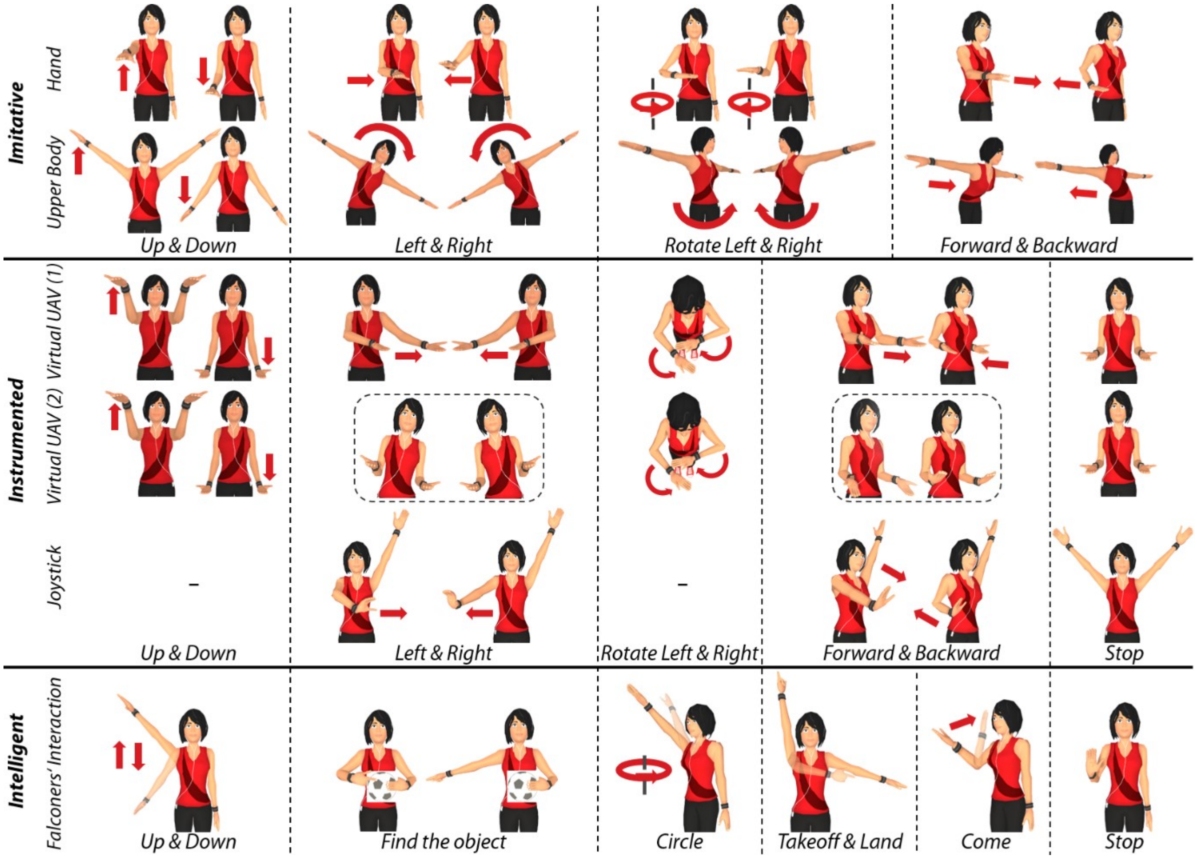


Figure 1. Examples of gesture sets (minor rows) for each of the three classes of mental models (major rows) for the specified commands (columns).

Selecting input commands according to a single metaphor promises to promote intuitive interaction. A system is considered to be *intuitive* if the way it works corresponds to our expectations. Thus, it should be fast to learn and easy to use. Mental models that define our expectations are formed by previously acquired knowledge and experiences. Through the use of metaphors that refer to common knowledge, it is possible to evoke certain mental models and, thus, encourage certain behavioral patterns. We consider a gesture set *intuitive* if a single hint is enough to define all its gestures. In order to evoke a certain mental model and, as a result, a certain behavior, a user needs to understand the metaphor that guides to the intended mental model.

Previous studies on natural interaction with UAVs have used various input vocabularies. The key distinctive feature of the suggested vocabularies is the underlying mental model that is used to draw an analogy between an input command and a vehicle motion. Prior to presenting a new classification scheme, we provide an example that demonstrates what is meant by a mental model in this context. Suppose a person is asked to navigate a UAV along a certain path using gestures. Most likely, the first question that would arise is 'Which gestures should I use?'. Then, instead of acquainting the person with a vocabulary, he is told to imagine holding a virtual UAV. Steering a UAV could then be realized by simply mapping the motion of the virtual UAV to the real one (Figure 1, *Virtual UAV 1-2*). Following this hint, a person can intuitively control the flight of the UAV without further instructions. In this case, the given scenario suggests a certain behavioral pattern. Intuitive interaction is achieved by giving hints or, in other words, by priming the users' mind with particular ideas. These ideas associated with different behaviors are related to certain mental models. Therefore, a mental model defines behavior that becomes intuitive for an individual or a group of individuals under a certain scenario.

Many of the works discussed below try to achieve intuitiveness of interaction by making use of different metaphors that evoke certain mental models. We propose the following classification scheme of mental models: *Imitative*, *Instrumented*, and *Intelligent*. In the *Imitative* class, a direct mapping between a performed gesture and a vehicle motion is used, e.g., rotation of the head changes the vehicle's orientation. The *Instrumented* class is defined by the presence of an illusion about navigating a vehicle through an intermediate link, e.g., a joystick. Associating a vehicle with a certain living creature that is equipped with some sort of intelligence is a key feature of the *Intelligent* class, e.g., treating a UAV as if it were a bird.

Next, the presented classes are described in detail. Underlying mental models of the reported input vocabularies are discussed along with the analysis of these vocabularies with respect to intuitiveness. The concept of mental models is used to answer questions raised in Section 1.

2.1 Imitative Class

The *Imitative* class implies that a vehicle is capable of imitating movements performed by an operator. This interaction can be seen as a direct mapping of operator's movements to the vehicle motion. In order to exemplify this idea, several input vocabularies are presented. Among them are hand, head, upper body, and full body mental models.

Hand. Liebeskind [4] made a demonstration of navigating a UAV partially using the *Imitative* class of mental models. The author used a correspondence between hand and vehicle motion to command a UAV to move up, down, left, right, forward, and backward, whereas to command a UAV to take off and land a user had to perform a gesture that mimics a 'left-click' gesture familiar to computer users. However, this gesture belongs rather to the *Instrumented* class (discussed in Subsection 2.2) as the intermediary link in a form of the imaginary mouse is used. Therefore, this example uses a mixture of gestures from different classes of mental models. The hypothesis that switching between classes leads to a higher mental workload seems to be reasonable. However, a formal study to test this hypothesis is needed. Another interesting aspect is that in order to command a UAV to rotate left and right, an operator rotates the hand about the horizontal instead of the vertical axis. This choice can be explained by the fact that the expected gestures (Figure 1, *Hand*) are not physically ergonomic.

Head. Higuchi and Rekimoto [5] suggested a gesture set based on the idea of synchronizing the position and orientation of an operator's head with those of a UAV. This gesture set is considered to be intuitive as its gestures belong to the single mental model and, therefore, could be easily guessed by novice users. The limitation of the presented vocabulary is that an operator's workspace has to be of the same size as the environment to explore. To overcome this limitation, the authors suggested commanding a UAV to fly forward and backward by tilting one's head forward and backward, respectively.

Upper Body and Full Body. Pfeil et al. [7] presented the following example of the upper body *Imitative* class: An operator navigates a UAV with arms spread to imitate wings (Figure 1, *Upper Body*). The employed mental model implies that the operator bends or rotates the upper body depending on the desired vehicle movement. However, the *up* and *down* gestures do not perfectly fit to the underlying mental model. Ideally, the operator would try to move higher by standing on toes to command a UAV to go up and move lower by bending knees to command it to go down. However, the final selection of the gestures might have been influenced by the aspects related to physical ergonomics.

Pittman and LaViola [6] presented an example of the full body *Imitative* class. In order to command a UAV to fly left, right, forward, and backward, an operator simply takes a step to the corresponding side. For the rotation commands, the operator rotates to the required side. Standing on toes and squatting down is used to command a UAV to fly up and down.

2.2 Instrumented Class

The *Instrumented* class suggests that an operator controls a vehicle through an imaginary intermediate link that can be an imaginary *physical object*, e.g., a joystick, a *link* that allows to manipulate a vehicle like a marionette or the ability to use *super force* to move a vehicle without touching it, e.g., repelling or attracting a vehicle with an open palm. Interaction techniques related to this class exploit the operator's assumptions that are based on knowledge and experience about certain objects or activities. For example, when controlling the flight of a UAV using an illusion of doing it with a joystick, an operator needs prior knowledge about the way the device works.

The literature review has revealed a few examples that exploit the *Instrumented* class. Wheller [20] presented a virtual joystick and keyboard interfaces used to control the flight of a simulated UAV. Another example of navigating an Unmanned Ground Vehicle (UGV) with an imaginary joystick was presented by Fong et al. [21]. The suggested vocabulary is shown in Figure 1 (*Joystick*). The raised left arm indicates the gesture-based interaction mode, while the right hand specifies a direction to move. Provided that an operator has the knowledge and experience needed to use a joystick, all the gestures are intuitive besides the *stop* gesture that represents an outlier. In addition, the requirement to keep the left arm up reduces intuitiveness and increases physical demand of the gesture set.

Two other examples of the *Instrumented* class are presented by Pfeil et al. [7]. These gesture sets also exploit the idea of using an imaginary *physical object* to navigate a UAV. In particular, an operator imagines holding a virtual UAV. While keeping this idea in mind, the operator changes the position and orientation of this UAV. These changes are directly translated into movements of the real UAV. As soon as the operator returns hands to the neutral position, a UAV stops its current movement. In the first gesture set (Figure 1, *Virtual UAV 1*), the operator is standing while steering a UAV, while in the second gesture set the operator is sitting. Apart from the positions, different gestures are used to command a UAV to fly to the left and right sides. In the second gesture set, instead of moving both hands to the corresponding side, the operator moves the left hand slightly down and the right hand slightly up. These gestures might be more intuitive for users with an experience in steering a UAV who are aware that in order to move a UAV forward, backward, left, and right it is required to tilt it forward, backward, left, and right, respectively. Following this logic, the gestures used in this set for the *forward* and *backward* commands should be as shown in Figure 1 (*Virtual UAV 2*) to constitute a coherent set. However, possibly due to implementation-related problems (small amplitude movements are harder to recognize), the authors employed the gestures from the first gesture set.

In their study, Pfeil et al. examined six different input vocabularies. Among them are (1) the original touchscreen-based interface, (2) the vocabulary associated with the *Imitative* class (Figure 1, *Upper Body*), (3-4) two vocabularies associated with the *Instrumented* class described here, (5) the vocabulary based on the assumption that an operator is a king/queen, and (6) the vocabulary that implies that the hands of an operator represent the control sticks of a typical game controller. Gestures used in (5) and (6) are unlikely to be intuitive for a novice user as there are no obvious mental models associated with the given scenarios. This statement is supported by the analysis reported in the paper: the majority of participants evaluated vocabularies (1), (5), (6) as the least natural and vocabularies (2)-(4) as the most natural. This result underlines the importance of using mental models when developing a vocabulary.

2.3 Intelligent Class

The key feature of the *Intelligent* class as its name implies is that a UAV is treated as an intelligent creature. This explains the fact that, in many cases, this class is deemed to resemble natural interaction the most. For example, when a person is instructed where to go, people tend to describe a place verbally and redundantly point out a direction with their index finger. This kind of interaction was observed by several researchers who explored the natural behavior of novice users when steering a single UAV [15,16] and multiple UAVs [14].

Recently, Lichtenstern et al. [22] proposed an interaction technique that makes use of a pointing gesture to activate one particular UAV out of a group of UAVs. By selecting UAVs one by one, an operator composes a team. As soon as the team is selected, the operator lifts the left arm. Next, the UAVs are imitating movements shown by the operator's right hand. It is interesting to note that the described technique exploits a combination of the *Intelligent* and *Imitative* classes. The former is used to activate an agent, while the latter is used to control movements of the selected vehicle(s). However, the intuitiveness of the interaction is reduced by the following two aspects: (1) each time a vehicle is selected, the operator confirms the selection by touching the right arm with the left hand; (2) the operator has to lift the left arm to indicate a switch from the 'selection' to the 'navigation' phase. The use of 'out of mental model' gestures might be the main source of errors observed during training sessions. The validity of this statement would be interesting to test in further work. Nagi et al. [10] also suggested using a pointing gesture to select an individual vehicle. A two-handed pointing gesture is used to select a group of UAVs that fall within the indicated range, and to select all the vehicles, an operator puts her hands together.

Eye contact is another important non-verbal communication channel that is a vital component of face-to-face communication. Often it is sufficient to look at a person to attract his attention. This technique that is characteristic to Human-Human interaction has inspired several research works. Couture-Beil et al. [23] presented a technique where an operator simply looks at a vehicle to activate it. Milligan et al. [9] suggested a technique that allows interacting not only with a single vehicle but with a group of vehicles. A vehicle or a group of neighboring vehicles is selected when an operator looks at it while making a gesture 'encircling' it. Then, the operator shows a target location using the pointing gesture. In contrast to the technique described previously, the *Intelligent* class is used for both, to select and navigate vehicles.

As an extension of previous works to the 3D case, Monajjemi et al. [12] presented a technique of interacting with a team of UAVs. Similarly, an operator selects a UAV by looking at it. However, in this case, the operator can add or remove a UAV to or from a selected group of UAVs using the following vocabulary: the right- or left-hand wave corresponds to adding or removing a vehicle, respectively, the right- and left-hand wave activates all UAVs. In terms of intuitiveness, it can be noted that a gesture used to select a UAV might be associated with the 'hello' gesture. Just as the hand wave gesture can be used to attract the attention of a person, the operator performs this gesture to attract the UAV's 'attention'. However, the left-hand wave and both hand-wave gestures are not that intuitive due to the lack of obvious association with corresponding commands, but at least they can be learned easily.

Ng and Sharlin [3] associated the flight of a UAV with the flight of a bird and presented a gesture set inspired by falconers' interaction with birds (Figure 1, *Falconer's Interaction*). Admittedly, this gesture set is intuitive for a specific group of people, but not for novice users. The authors asked several participants to test and give their feedback on the suggested gesture set. Unsurprisingly, the gestures for the *stop* and *come* commands were selected as the most intuitive. Obviously, the reason why participants considered these gestures as intuitive is their frequent use in day-to-day communication.

It is worth mentioning that users tend to give high-level commands when using the *Intelligent* class. For example, to guide a vehicle to a target point, a user would just point out a direction instead of giving low-level commands, e.g., up, left, and forward. This can be explained by the fact that a user associates a vehicle with an intelligent creature, which in turn leads to higher expectations about its capabilities.

2.4 Discussion

The key difference of the presented classes is the expectations they raise. The highest expectations are associated with the *Intelligent* class. In this case, an operator assumes that a vehicle is able to interpret high-level commands. The *Instrumented* class implies that a vehicle is able to translate given low-level commands. The lowest expectations correspond to the *Imitative* class, where a vehicle simply copies the gestures. The *Intelligent* and *Instrumented* classes require a more complex interpretation mechanism compared to the *Imitative* class where it is enough to track operator's body movements and directly map them to the movements of a UAV. The need for initial instructions is another distinctive feature of the classes. Ideally, techniques associated with the *Intelligent* class allow an operator to navigate a system following the natural 'flair' without a need for prior instructions. For the *Imitative* and *Instrumented* class, a hint specifying the type of interaction is needed. Besides the hint, the *Instrumented* class requires certain knowledge and experience from an operator. It seems that the *Intelligent* class should be given a preference. However, in some cases operators might feel unnatural to interact with a vehicle as with an intelligent creature. A study is needed to resolve this issue.

The concept of mental models is a powerful tool that helps to develop an intuitive interaction technique. Interaction designers must consider various aspects when choosing a proper mental model. The target group is among the key aspects that have to be considered when defining a vocabulary. As mentioned previously, while the vocabulary based on the falconer's interaction may be intuitive for a specific group of people, it is not for novice users. Thus, in order to achieve intuitive interaction with a system, the key requirement is the involvement of a focus group during the development of the input vocabulary in order to observe natural behavior. From this behavior the mental models can be derived, which potentially guided users in their choices for commands. The identified users' mental models should be considered when defining an input vocabulary with a preference given to the most frequently observed models. Evoking a certain mental model helps to 'understand' the entire vocabulary instead of memorizing each gesture individually. Thus, we recommend to avoid, if possible, a mixture of mental models for commands associated with a common type of tasks (e.g., navigation tasks or formation control) as switching between different mental models might lead to a higher level of mental workload and cause errors due to confusion of commands.

In addition, physiological (e.g., left- and right-handedness) and cultural differences (e.g., nodding head means 'yes' or 'no', depending on the society) have to be taken into account. Important aspects related to physical ergonomics do not need to be explained. The field of application is another aspect to consider, as it has a significant impact on interface requirements and on interaction techniques accordingly. For example, a technique that requires intensive physical effort would have the potential to be applied in an entertainment sector, e.g., video games, but it might not be acceptable for serious applications, e.g., search and rescue missions.

To sum up, in order to achieve intuitive interaction, (1) the input vocabulary has to employ mental models that are known to the considered group of people; (2) the natural behavior of users has to be analyzed to discover user-defined mental models; (3) mixing gestures from different mental models in an input vocabulary should be avoided; (4) important aspects such as physiological and cultural differences, physical ergonomics, and the field of application have to be considered in the final input vocabulary.

The works discussed here focus on users with little or no prior experience with UAVs. Thus, the suggested input vocabularies cover only commands for basic interaction with UAVs (e.g., navigation or forming a group of UAVs) and are not intended to replace complex GCSs and WIMP interfaces. As we have seen, more advanced commands such as *takeoff* and *land* are left out by many of the discussed works, while the basic motion commands as *up*, *down*, *left*, *right*, *rotate left*, *rotate right*, *forward*, and *backward* are covered by most researchers. One reason might be that it is difficult to find a gesture that would fit well. In such cases, the natural limit of expressive power of gestures can be compensated with accompanying voice commands.

Previous work on natural interaction with UAVs has shown that users, in case they had freedom to choose, tended to mix gestures from different mental models [16]. The inability of one mental model to cover all the commands (e.g., the 'left-click' gesture discussed in Subsection 2.1) and factors related to physical ergonomics (e.g., the *Up & Down* gestures in Figure 1, *Upper Body*) are among possible reasons that induce users to mix mental models. Admittedly, these issues are limitations of 'single mental model' vocabularies. Another reason is that it might be simply more natural for users to randomly switch between different mental models rather than to stick to a single one when in 'gesture storming' mode. Ideally, an interface has to allow users to interact with a system as they prefer by letting them spontaneously switch between mental models. A vocabulary of such an interface has to include several vocabulary entries for the considered commands. Extension of the input vocabulary with all the possible 'synonyms', if at all possible, is most likely to significantly complicate implementation of the system. In addition, developers have to consider cultural differences that might cause different interpretations of some vocabulary entries and deal with homonyms (when the same entry is used for different commands). Thus, a pragmatic approach is to limit the users' 'freedom' of choice for the input vocabulary by guiding them to a certain mental model that uniquely defines the way to interact with a system.

The overview of interaction techniques revealed that integration of multiple modalities is still in its infancy and further research is required. A combination of natural modalities such as gestures, speech, gaze direction, and facial expressions has the potential to further increase the naturalness and intuitiveness. Previous work has demonstrated strong advantages of interfaces that allow a user to interact with a system through multiple 'natural' modalities [30]. First steps have been taken to explore a combination of gestures with speech for UAV navigation [14,15,16]. It shows that instead of sticking to one modality, the users tend to combine speech and gestures. This finding confirms that a multimodal interaction is indeed more natural. Further work is needed to investigate other natural modalities and their combinations. The art of developing intuitive techniques for multi-agent UAV systems is another area that requires exploration. The discussed techniques consider the case of navigating a group of UAVs as a single agent meaning that once a group of UAVs is selected, all group members perform identical actions. Further research is needed to develop intuitive and natural interaction techniques for multi-agent UAV systems that include specific commands for group interaction such as *split*, *get together*, and commands for formation control.

Most of the techniques assume that a UAV system is within operator's field of view. In this case, the use of natural input modalities is likely to increase the naturalness and intuitiveness of Human-UAV Interaction. Flight navigation with indirect observation of the system requires an in-depth study. Another aspect that is still in its infancy is 'socialization' of UAVs that defines the expected and acceptable behavior of a UAV in a certain scenario. Currently, we observe first attempts to understand how users envision a UAV as a 'companion' [31,32].

3 Conclusion

It is a challenging and multidisciplinary problem to develop natural and intuitive interaction techniques for UAV systems. This article provided an overview of the work done so far, introduced a classification scheme, and analyzed the interaction techniques in terms of intuitiveness. In the scope of our work, we defined a notion of intuitiveness as a feature of an input vocabulary that makes all its entries apparent for users with or even without a single hint about the underlying mental model. We strongly argue for considering the underlying mental models when developing an interaction technique since it is a key aspect when defining the vocabulary and when further testing and analyzing the intuitiveness.

We introduced a classification scheme to group input vocabularies for UAV systems based on their underlying mental models. Using this scheme, we clustered the discussed techniques into three classes – *Imitative*, *Instrumented*, and *Intelligent*. Each class has been defined and illustrated with corresponding techniques. The introduced concept was used to assess the intuitiveness of the interaction techniques. Table 1 provides references in accordance with the introduced classification scheme.

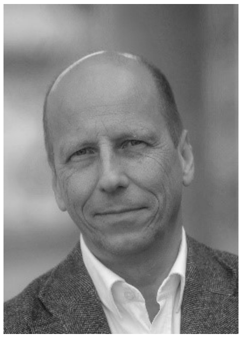
Table 1. Classification scheme of input vocabularies for UAV systems.

Scope	*Imitative*	*Instrumented*	*Intelligent*
A single UAV	[4],[5],[6],[7]	[7],[20]	[3],[11],[15]
A group of UAVs	-	-	[10],[12],[14],[22]

The literature overview has revealed several gaps that should be covered: (1) multi-modal interaction and (2) the potential of gaze direction, facial expressions, and speech as input modalities are still laid aside and further research is needed; (3) most of the presented works are focused on single agent UAV systems and there are just a few papers addressing more complex multi-agent systems; (4) the majority of gesture-based interfaces are considered for cases when a UAV system is within operator's field of view and it remains open whether gesture-based interaction would be beneficial in case of indirect observation; (5) various aspects related to 'socialization' of UAVs have to be further investigated.

The Authors

Ekaterina Peshkova is a PhD candidate within the Joint Doctoral Programme in Interactive and Cognitive Environments in the Department of Informatics Systems at Alpen-Adria-Universität Klagenfurt and the Electrical, Electronics and Telecommunication Engineering and Naval Architecture Department at Università degli Studi di Genova. Her research interests include HCI and user-centered design. Peshkova received an MSc in Robotics (double degree) within the EMARO (European Master on Advanced Robotics) programme from École Centrale de Nantes and Università degli Studi di Genova. Contact her at ekaterina.peshkova@aau.at



Martin Hitz is a professor of Interactive Systems at Alpen-Adria-Universität Klagenfurt, Austria, and currently also vice rector of the university. His research interests include HCI, usability, and non-standard user interfaces, generally taking into account also mobile systems and pervasive interfaces. Prof. Hitz received his PhD in computer science from the Technical University of Vienna and his habilitation (venia docendi) from the University of Vienna. He is a member of ACM and ACM SIGCHI. Contact him at martin.hitz@aau.at

Bonifaz Kaufmann is CTO and Co-Founder of the Internet of Things 40 Systems GmbH in Austria. Previously he worked in management positions in R&D, Product Management and as a technology consultant. His research is at the intersection of HCI and the Internet of Things (IoT). He has studied Informatics at the University of Klagenfurt, the University of Westminster in London and at the MIT Media Lab in Cambridge, USA. He received his PhD in Technical Sciences for his research contributions in HCI and Handheld Projector Interaction. He is the creator of the Amarino Toolkit, one of the first available IoT prototyping kits for connecting mobile phones and microcontrollers. Contact him at bonifaz.kaufmann@gmail.com.

References

1. Kevin W. Williams. 2004. A summary of unmanned aircraft accident/incident data: Human factors implications. In *Final Report DOT/FAA/AM-04/24*, US Department of Transportation, Federal Aviation Administration, (December 2004).
2. Florian ‘Floyd’ Mueller and Matthew Muirhead. 2015. Jogging with a quadcopter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 2023-2032.
3. Wai Shan Ng and Ehud Sharlin. 2011. Collocated interaction with flying robots. In *Proceedings of the 20th IEEE International Symposium on Robot and Human Interactive Communication* (RO-MAN '11), 143-149.
4. Daniel Liebeskind. (2013, November 19). Leap Motion Node Drone Flight in 2013 [Video file]. Retrieved from https://www.youtube.com/watch?v=hfq2SisPvCU.
5. Keita Higuchi and Jun Rekimoto. 2013. Flying head: A head motion synchronization mechanism for unmanned aerial vehicle control. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 2029-2038.

6. Corey Pittman and Joseph LaViola. 2014. Exploring head tracked head mounted displays for first person robot teleoperation. In *Proceedings of the 19th International Conference on Intelligent User Interfaces* (IUI '14), 323-328.

7. Kevin Pfeil, Seng Lee Koh, and Joseph LaViola. 2013. Exploring 3d gesture metaphors for interaction with unmanned aerial vehicles. In *Proceedings of the International Conference on Intelligent User Interfaces* (IUI '13), 257-266.

8. John Paulin Hansen, Alexandre Alapetite, I. Scott MacKenzie, and Emilie Møllenbach. 2014. The use of gaze to control drones. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (ETRA '14), 27-34.

9. Brian Milligan, Greg Mori, and Richard T. Vaughan. 2011. Selecting and commanding groups in a multi-robot vision based system. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (HRI '11), 415-416.

10. Jawad Nagi, Alessandro Giusti, Luca M. Gambardella, and Gianni A. Di Caro. 2014. Human-swarm interaction using spatial gestures. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS '14), 3834-3841.

11. Jawad Nagi, Alessandro Giusti, Gianni A. Di Caro, and Luca M. Gambardella. 2014. Human control of UAVs using face pose estimates and hand gestures. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (HRI '14), 252-253.

12. Valiallah (Mani) Monajjemi, Jens Wawerla, Richard Vaughan, and Greg Mori. 2013. HRI in the sky: Creating and commanding teams of UAVs with a vision-mediated gestural interface. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS '13), 617-623.

13. Nataliya Kos'myna, Franck Tarpin-Bernard, and Bertrand Rivet. 2014. Bidirectional feedback in motor imagery BCIs: Learn to control a drone within 5 minutes. In *CHI EA '14 Extended Abstracts on Human Factors in Computing Systems*, 479-482.

14. Geraint Jones, Nadia Berthouze, Roman Bielski, and Simon Julier. 2010. Towards a situated, multimodal interface for multiple UAV control. In *Proceedings of the IEEE International Conference on Robotics and Automation* (ICRA '10), 1739-1744.

15. Jessica R. Cauchard, Jane L. E, Kevin Y. Zhai, and James A. Landay. 2015. Drone & me: An exploration into natural human-drone interaction. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (UbiComp '15), 361-365.

16. Ekaterina Peshkova, Martin Hitz, and David Ahlström. 2016. Exploring user-defined gestures and voice commands to control an unmanned aerial vehicle. In *Proceedings of the 8th International Conference on Intelligent Technologies for Interactive Entertainment* (INTETAIN '16), [ACCEPTED].

17. Michael Burke and Joan Lasenby. 2015. Pantomimic gestures for human-robot interaction. *IEEE Transactions on Robotics*. 31, 5 (October 2015), 1225-1237.

18. Thi Thanh Mai Nguyen, Ngoc Hai Pham, Van Thai Dong, Viet Son Nguyen, and Thi Thanh Hai Tran. 2011. A fully automatic hand gesture recognition system for human-robot interaction. In *Proceedings of the 2nd Symposium on Information and Communication Technology* (SoICT '11), 112-119.

19. Jacob O. Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A. Myers. 2005. Maximizing the Guessability of Symbolic Input. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 1869-1872.

20. Kevin R. Wheeler. 2003. Device control using gestures sensed from EMG. In *Proceedings of the IEEE International Workshop on Soft Computing in Industrial Applications* (SMCia '03), 21-26.

21. Terrence W. Fong, Chuck Thorpe, and Charles Baur. 2001. Advanced interfaces for vehicle teleoperation: Collaborative control, sensor fusion displays, and remote driving tools. *Autonomous Robots*. 11, 1 (July 2001), 77-85.

22. Michael Lichtenstern, Martin Frassl, Bernhard Perun, and Michael Angermann. 2012. A prototyping environment for interaction between a human and a robotic multi-agent system. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (HRI '12), 185-186.

23. Alex Couture-Beil, Richard T. Vaughan, and Greg Mori. 2010. Selecting and commanding individual robots in a vision-based multi-robot system. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (HRI '10), 355-356.

24. David Efron. 1941. Gesture and environment. Morningside Heights, New York: King's Crown Press.

25. David McNeil. 1992. Hand and mind: What gestures reveal about thought. The University of Chicago Press.

26. Paul Ekman. 1999. Emotional and conversational nonverbal signals. In Messing, L. S. & Campbell, R. (Eds.), *Gesture, Speech, and Sign*, 45-55.

27. Adam Kendon. 1988. How gestures can become like words. In F. Poyatos, ed., *Crosscultural Perspectives in Nonverbal Communication*. Toronto: C. J. Hogrefe, Publishers, 131-141.

28. Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09) 1083-1092.

29. Jaime Ruiz, Yang Li, and Edward Lank. 20011. User-defined motion gestures for mobile interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 197-206.

30. Rajeev Sharma, Vladimir I. Pavlović, and Thomas S. Huang. 1998. Toward multimodal human-computer interface. In *Proceedings of the IEEE*. 86, 5 (May 1998), 853-869.

31. Hyun Young Kim, Bomyeong Kim, and Jinwoo Kim. 2016. The naughty drone: A qualitative research on drone as companion device. In *Proceeding of the 10th International Conference on Ubiquitous Information Management and Communication* (IMCOM '16), no. 91.

32. Jessica R. Cauchard, Kevin Y. Zhai, Marco Spadafora, and James A. Landay. 2016. Emotion encoding in human-drone interaction. In *Proceeding of the 11th ACM/IEEE International Conference on Human Robot Interaction* (HRI '16), 263-270.