Link to publisher version with DOI

10.1109/THMS.2019.2919719

A User-Centered Framework for Designing Mid-Air Gesture Interfaces

Antonio E. Uva, Michele Fiorentino, Vito M. Manghisi, Antonio Boccaccio, Saverio Debernardis, Michele Gattullo, Giuseppe Monno, *DMMM, Polytechnic University of Bari, Bari, Italy*

***Abstract*— Due to the recent advances in technologies for gesture-recognition, mid-air gestures can be considered the interface of the future in a large number of applications. However, designing effective interfaces with mid-air gestures is not an easy task because the design is application dependent and it must fulfill many requirements at the same time. Despite the availability of general guidelines in the literature, clear and well-established procedures for the optimal design of mid-air gesture-based interfaces are, to date, not available and remain an open issue. The main contribution of this work is a user-centered modular framework, which integrates existing and novel methods. It supports the designer considering multiple aspects including ergonomics, memorability, and specific user requirements tailored to the application scenario. The framework involves three design steps and a final validation step, also supported by dedicated software. We tested with success the proposed framework in an industrial case study, where technicians must easily access technical information by browsing digital manuals during maintenance operations.**

***Index Terms*— Gesture vocabulary, mid-air gesture interface, user-centered elicitation approach, ergonomics, consumed endurance.**

I. Introduction

Mid-air gestures have a still unexplored potential in a wide number of application fields [1], [2]. One of the main advantages of mid-air gestures is that they can provide natural and intuitive interactions with a lower cognitive effort compared to traditional interfaces [3]. Another big advantage of mid-air gestures is that they do not require the user to wear or handle a tracked device. Thus, contrarily to wearable devices (e.g., data gloves [4]), they do not interfere with workwear (e.g., sterilized gloves by surgeons, anti-hardship gloves by workers). Moreover, contrarily to handheld devices, they leave hands free for working activities.

Compared with other input methods with the same advantage, such as speech-based ones [5], mid-air gestures can also be used in noisy environments (e.g., an industry) and where, conversely, silence must be respected (e.g., a museum).

Furthermore, mid-air gestures are also suited for Augmented and Virtual reality applications and may improve the sense of immersion.

Recent advances in computer vision, hand, and body tracking algorithms allow real-time gesture recognition [1], [2], [6], [7]. However, designing effective interfaces that leverage mid-air gestures, is not an easy task because it is application dependent and must fulfill many requirements. For example, gestures must be: easy to perform and remember, intuitive, metaphorically connected with their functionality, and not physical demanding [8]. Gestural communication may involve a muscular effort greater than traditional input or speech. In particular, long sessions of mid-air interactions may lead to upper limbs fatigue, a condition known as “the gorilla arm effect” [9].

Many studies in the literature propose a user-centered approach based on gesture elicitation, where users are asked to suggest intuitive gestures for each command. However, those gestures though intuitive may not fulfill other requirements. Despite the availability of general design guidelines in the literature [1], [10], clear and well-established procedures for the optimal design of mid-air gesture-based interfaces are, to date, not available and remain an open issue.

The main contribution of this work is a user-centered framework, which complementary integrates existing procedures/methods with novel elements. It supports the mid-air interface designer considering multiple aspects including ergonomics, memorability, and specific user requirements tailored to the application scenario. The framework involves three design steps and a final validation step, supported by a dedicated software.

We tested the proposed framework with a case study, which our research group is familiar with [7], [11], [12]. We choose an industrial maintenance scenario, where technicians must easily access technical information during maintenance operations, by browsing digital-manuals.

II. Related works

The design of an optimal gesture-based control vocabulary requires a tradeoff among various factors: the accuracy and the speed of recognition, the intuitiveness, and the ergonomics of the gesture. Stern *et al.* [13] distinguished three main approaches for gesture vocabulary design: (i) the centrist or authoritarian approach, where the designer decides which vocabulary should be used [14]; (ii) the user-centric or consensus approach, where a group of users decides on a

Manuscript received

Authors are with the Department of Mechanics, Mathematics and Management, Polytechnic Institute of Bari, Italy. (e-mail {antonio.uva, michele.fiorentino, vitomodesto.manghisi, antonio.boccaccio, saverio.debernardis, michele.gattullo, giuseppe.monno}@poliba.it)

common vocabulary to express a given set of commands [15]; (iii) the individual or customized approach, where each individual defines his/her own vocabulary [16].

The user-centric approach is commonly preferred to other approaches [8], [17–19] and represents the standard methodology in natural user interfaces. This approach was successfully adopted by Piumsomboon *et al.* [20] for augmented reality applications. They elicited user-defined gestures and observed that users tended to adopt reversible gestures (i.e., gestures performed in opposite directions for commands with opposite effects).

A crucial important objective of the user-centric methodology is lowering of the cognitive effort and improving the user experience. Wobbrock *et al.* [21] proposed a unified approach to evaluate and maximize the intuitiveness of symbolic inputs and formalized two metrics: (i) the "*guessability*" - intuitiveness of a specific gesture compared to others - and (ii) the "*agreement*" - the consensus of a specific set of commands among users-. They also addressed the conflict set problem, i.e., the same gestures are proposed for more than one command.

To consider the intuitiveness of the gestures, the speed, and accuracy of the recognition system, measurements of technical factors were carried out. Stern *et al.* proposed a methodology that considered both psycho-physiological measures (intuitiveness, comfort) and gesture recognition accuracy [22], [23]. Their results, obtained with static gestures, can be used as a data repository of intuitiveness and comfort measures. Hessam *et al.* [24] focused on gesture set optimization by assessing and minimizing possible confusions deriving from the use of the same gesture for different commands. Pereira *et al.* [25] took into account user ratings, cognitive load and ergonomics. They created a 3-D hand gesture set for common HCI (human-computer interactions) tasks guided by user-generated gestures, with the final selection based on user ratings, estimation of postural risk, and consideration of system capabilities.

The user-centric approach presents two potential pitfalls: the *legacy bias* [26], [27], and the *performance bias* [28]. The *legacy bias* occurs because users are often biased by their prior habits and legacy technologies that have been used for a long time. The *performance bias* because, in the elicitation phase, the aspects related to the repetitiveness of the gesture and hence to the consequent fatigue are not considered. Possible solutions to overcome these potential pitfalls were recently proposed [26] [28]. Two main approaches are currently utilized to evaluate users' fatigue. The first one employs subjective ratings acquired through interviews or questionnaires, such as the CR10 Borg scale [29], and the NASA-TLX [30], [31]. The second approach -objective- is based on direct measurements such as intramuscular pressure and tissue oxygenation [32], electromyography [33] and limb position [34], [35]. Another quantitative approach was recently proposed by Hincapié-Ramos *et al.* [36]. They proposed a new metric to assess arm fatigue by the so-called consumed endurance (CE). CE, - which was adopted in this study -, is based on Rohmert's formulation [37] that expresses the maximum amount of time that a muscle

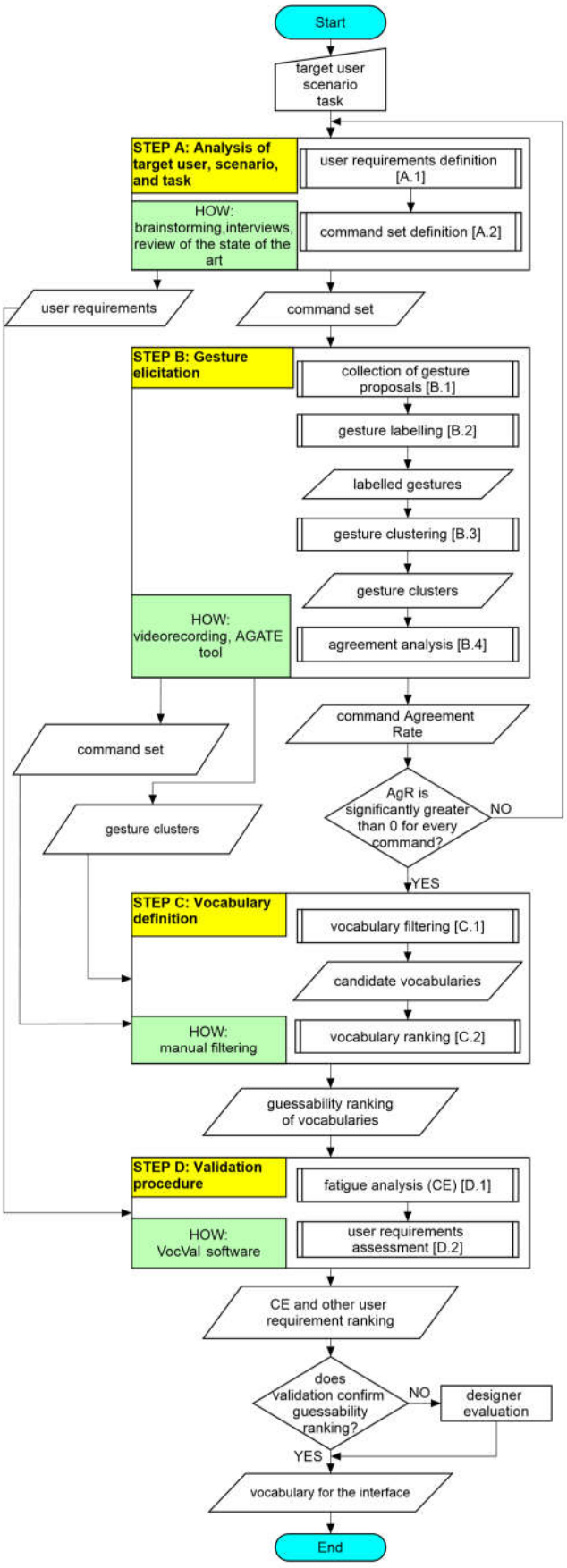

Fig. 1. Flow-chart of the proposed framework.

can maintain a contraction level before needing rest. Recently the American Conference of Governmental Industrial Hygienists [38] formalized an exponential expression for the upper-limb exertion time-limit fatigue.

By reviewing the state-of-the-art, we can conclude that the user-centric approach has been used for the designing of gesture-based vocabularies. However, it should be integrated with other measurements to address issues such as fatigue, memorability [8], and legacy and performance bias. We set up a general framework to design mid-air gesture-based vocabularies with a user-centered approach.

III. The Proposed Framework

The following definitions are extensively used in the description of the proposed framework:

- *Command:* a functionality to be triggered by the user using the interface;
- *Task:* the goal to be achieved by the user using a command-set;
- *Gesture:* the combination of user movements and postures aimed at triggering a single command.
- *Vocabulary*: a mapping of gestures to the command-set;

Based on general requirements suggested by Blake [39], in a preliminary phase we have defined the following interface requirements (IR):

IR 1 The interface must execute a set of commands, to perform a specific task as required by the scenario;
IR 2 The candidate vocabularies consist of gestures proposed by users;
IR 3 There must be an adequate level of consensus among users for all the commands;
IR 4 The selected vocabulary must be coherent: without conflicts, homogeneous, and reversible;
IR 5 The selected vocabulary must be intuitive;
IR 6 The gestures composing the selected vocabulary must imply low physical fatigue and high memorability;
IR 7 The selected vocabulary must respect specific user requirements defined according to the specific scenario and task.

To meet these requirements, the proposed framework provides a 4-step procedure (Fig. 1) where each step was designed to meet the IRs:

Step A: *analysis of target user, scenario, and task* (IR 1, IR 7).
Step B: *gesture elicitation* (IR 2, IR 3);
Step C: *vocabularies definition* (IR 4, IR 5);
Step D: *validation procedure* (IR 6, IR 7).

A. Analysis of target user, scenario, and task

The analysis of user, scenario, and task can be accomplished using brainstorming between the stakeholders (e.g., designers, end-users, and customers), interviews, and a review of the state of the art. The output is the definition of the user requirements (IR7), the expected users' capabilities and performance, considering ergonomic, cognitive, motivational, and technical factors. Thus, it is possible to define a command-set for the task (IR1).

B. Gesture elicitation

In elicitation studies, users are shown *referents* (an action's effects) and are asked to propose the corresponding *signs* (interactions that result in the given referent) [27], [40]. In our framework, the referents are the commands, while the signs are the gestures.

This phase includes the following sub-steps:

1) Collection of gesture proposals

Our framework uses a conscious bottom-up approach to collect a set of gesture proposals among users (IR2) as in Nielsen et al. [8].

The approach requires an introductory slideshow explaining the test execution modality and each referent. Then, users are asked to think aloud about the possible gestures they would propose for them.

The proposals are collected according to the following procedure. Users watch an automatic sequence of slides, each of them presenting one referent, and simultaneously have to execute any hand gesture that they consider the best for that referent. Participants stand at a proper distance from a screen displaying the slideshow; meanwhile, a camera records the gesture executions. The sequence is a fixed series of the set of referents, repeated several times, and mixed according to a Latin Square design. The presentation time is progressively reduced from 3 to 1 second.

2) Gestures labeling

We identified six general attributes to label the gesture.

(i) *Number of hands* used to perform the gesture.

(ii) *Mode of hands*: i.e., dominant, if the user utilizes, exclusively or mainly, the dominant hand; non-dominant in the opposite case; specular if both the hands are utilized and moved symmetrically with respect to the middle sagittal plane.

(iii) *Direction of motion*: right to left or left to right mainly perpendicularly to the sagittal plane, forward or backward mainly perpendicularly to the coronal plane, top-down or bottom-up in the vertical direction), diagonal.

(iv) *Amplitude of the motion*: to distinguish between wide and narrow gestures an amplitude threshold A_t = 45 cm is fixed which is the mean anthropometric value of shoulder breadth

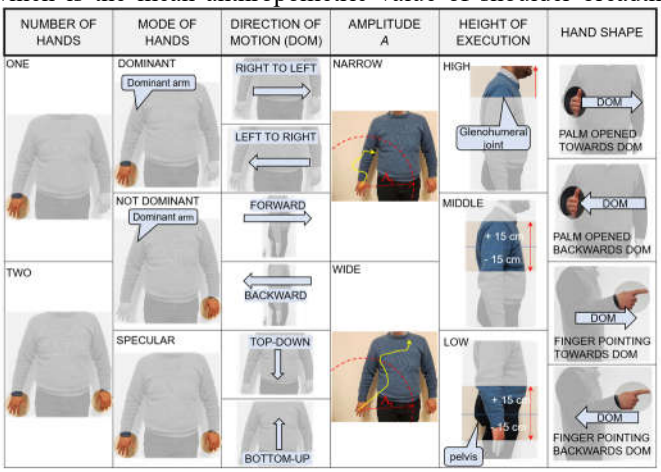

Fig. 2. The six general attributes utilized to label/classify (Step B.2) the collected gesture proposals and successively implemented in the case study.

[41]; every gesture with an amplitude A greater than A_t is classified as wide, and every gesture with $A < A_t$ is classified as narrow (Fig. 2).

(v) *Height of execution:* middle, if prevalently in a range of ±15 cm with respect to the glenohumeral joint height; high if above and low if below it.

(vi) *Hand shape*: e.g., fist closed, palm opened, finger pointing.

By watching the video recordings of gesture proposals, it is possible to label every gesture according to the previous attributes.

3) Gestures clustering

Labeled gestures are grouped in clusters using two criteria applied in sequence.

(i) Labeling-attribute significance: The information gain (IG), i.e., the expected reduction in entropy caused by partitioning the classes according to a given attribute, is evaluated for each labeling attribute. The higher the value of IG computed for a given attribute, the higher the capability of 'that' attribute of differentiating/distinguishing. The framework provides the assessment of the IG by implementing the supervised Information Gain Ranking Filter inside the WEKA tool [42].

(ii) Similarity rules: as suggested by Piumsomboon *et al.* [20], groups gestures according to similarities, that is by neglecting small differences in gesture execution.

4) Agreement analysis

This sub-step estimates the level of consensus (IR3) among gesture clusters for each referent using the *Agreement Rate (AgR)*. The *AgR* is defined as "the number of pairs of participants in agreement with each other divided by the total number of pairs of participants that could be in agreement" [43] and ranges from 0 – all proposals are different, to 1 – all proposals are the same.

AgR is evaluated with the AGATE tool [44] that also allows assessing the Variation between agreement rates statistics (V_{rd}). V_{rd} serves to verify if the *AgR* for every referent is significantly greater than 0. If this hypothesis does not hold, then the command-set should be thought again considering new interface requirements.

C. Vocabularies definition

This phase starts with the composition of all the possible vocabularies using all the combinations of gesture clusters for each command.

1) Vocabulary filtering

Interface requirement IR4 is applied as detailed in [10], [20], [21] to define candidate vocabularies for the task.

If within a vocabulary, two or more commands have the same gesture cluster (*conflict set*), the gesture is assigned exclusively to the command with the highest number of proposals.

Moreover, a vocabulary can be accepted as a candidate one only if its gesture clusters are *homogeneous*, and *reversible* for opposite commands.

2) Vocabulary ranking

The candidate vocabularies are ranked in terms of intuitiveness (IR5) evaluated using the *guessability G* [20], [21], [27], *G* is calculated as:

$$G = \frac{\sum_{k=1}^{N} |P_k|}{P_{TOT}} \; ; \; G \in]0, 1] \qquad (2)$$

where, N is the number of commands included in the vocabulary, $|P_k|$ is the number of gesture proposals in the cluster associated to the k^{th} command, and P_{TOT} is the total number of proposals collected in the elicitation phase. G ranges from 0 - gesture proposals are sparse, to 1 - all the users propose the same gesture for the same command.

D. Validation procedure

This phase validates the candidate vocabularies as regards fatigue and memorability (IR6), and other user requirements (IR7) defined in step A.

We designed a within-subject experiment to assess the metrics associated with fatigue (CE) and memorability (number of errors) and subjective evaluations.

Users stand in front of a monitor and a Microsoft Kinect device, at a distance of 200 centimeters.

In an initial training phase, a slideshow describes the gestures for the candidate vocabulary. Videos showing the gesture executions are embedded in the slideshow to explain the correspondence between gesture and command. Then, in the testing phase, participants read on a monitor (Fig. 3) a sequence of command names and execute the associated gesture in the vocabulary. Eventually, the experimenter marks an error (i.e., gesture missed-execution and mismatching) and the sequence of commands restarts from the beginning. The test finishes when the sequence is accomplished with no errors. The sequence is a fixed series of the command-set, repeated several times, and mixed according to a Latin Square design. To address possible learning effect, the order of execution is counterbalanced among participants with an adequate time interval and arranged in Latin Square design, too.

We developed the VocVal software, available at https://sourceforge.net/projects/mid-air-gesture-framework/ to support this validation step (Fig. 3). VocVal uses Kinect data and the CE workbench libraries [36] for the automatic calculation of the CE. Furthermore, the software allows the experimenter to mark users' errors. At the end of each test, the

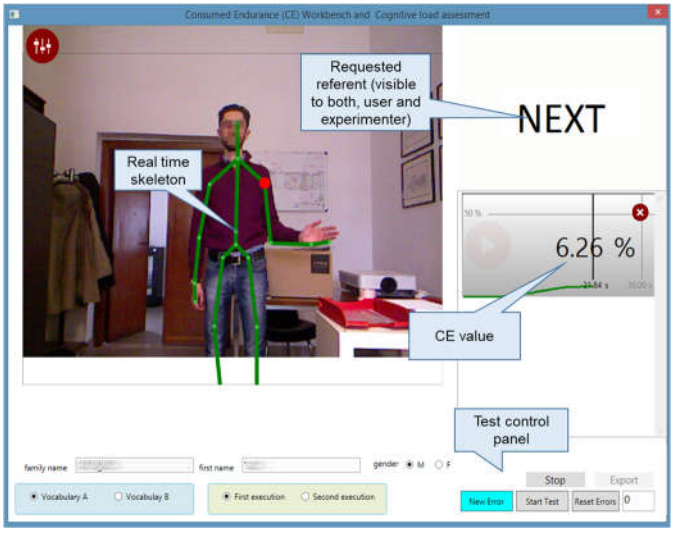

Fig. 3. Experimenter view of VocVal. Participants can see only the required command.

software allows administering the questionnaires for subjective evaluations.

The output of this step is a complete set of quantitative/qualitative indicators that guides the designer to choose the best vocabulary thoughtfully for specific applications.

IV. The Case Study

A. Analysis of target user, scenario, and task

In the industrial scenario, the task of browsing a digital maintenance manual is of utmost importance.

The analysis of this scenario and task was made through interviews with designers, customers, and maintenance operators along with a literature review. We found that instructions for a maintenance procedure in a digital manual can be successfully organized with a tree-like structure [11]. In this structure, a technician can go back and forth from an instruction to the previous/following one, skip well-known details and access to more specific information, if necessary. Thus, in our case study, we specify the task as the navigation of technical instructions using the following command-set:

- *Next.* A maintenance step is clear or completed, and hence the user wants to access the following information;
- *Previous.* The user wants to come back to the previous information;
- *Go down (to a lower level)*. More detailed information is required; therefore, the current step is expanded in a more detailed sequence of sub-steps;
- *Go up (to an upper level)*. The user needs fewer details; therefore, s/he navigates through a sequence of less detailed steps. Going up to the first level, the user reaches the root node of the manual.

In the navigation of a maintenance manual, operators should devote a low cognitive effort to remember the right gesture to execute, so that they can focus their attention on the task to accomplish rather than on the interface (IR6). The compliance with this user requirement can be measured through the memorability parameter.

In this case, we defined two further requirements (IR7) as (i) ease of execution and (ii) goodness of matching between commands and gestures. They both can be measured through Likert scale questionnaires. We require the median value of both the measures to be at least 5 on a 7-point scale.

B. Gestures Elicitation

1) Collection of gesture proposals

A population of 15 participants (average age 27.3 years, SD=5.11) was recruited (14 males and 1 female), all right-handed. All of them hold an MS in Mechanical Engineering; seven participants also hold a PhD in Mechanical Engineering. Three participants declared that they regularly used an Xbox Kinect for recreational purposes, nine utilized the device a few times, and three never used it.

Each referent, i.e., command, — *Next, Previous, Go down* and *Go up* — was clearly described; the participants were then asked to think aloud about the possible gesture proposals.

The gesture collection was performed for a fixed series of the four referents, repeated seven times (for a total of 7 × 4 = 28 gesture requests). The presentation time was progressively reduced according to the following sequence (number of referents × repetitions × presentation time): 4 × 3 × 3 + 4 × 3 × 2 + 4 × 1 × 1, for a total duration of 64 seconds.

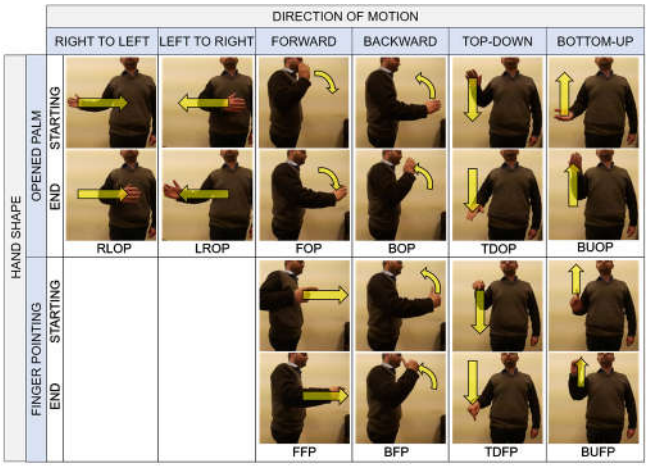

Fig. 4. Gesture clusters obtained combining the 2 most significant attributes and applying the similarity rules.

2) Gestures labeling

At the end of the tests, all videos were reviewed. Since some users missed executing the gesture as requested, only 397 proposals were collected instead of 28 requests × 15 participants = 420 as expected.

Labeled gestures were analyzed using descriptive statistics. Users preferred one-handed gestures (86.15%) compared to two-handed ones (13.85%). One-handed gestures, executed with the dominant hand, were mostly performed (89.17%), followed by specular gestures (10.33%). The direction of motion perpendicular to the transverse plane was preferred (50.12%), followed by the direction perpendicular to the sagittal plane (29.47%) and by the one perpendicular to the coronal plane (20.41%). Most users preferred wide movements (68.77%) rather than narrow ones (31.23%). As regards the height of execution, most users preferred middle (45.59%) and low gestures (43.83%) compared to high ones (10.58%). This can be explained by the argument that "high" gestures may involve shoulder abduction, thus resulting in awkward postures. We observed only two hand shapes among the possible ones. They are palm opened (77.08%) and finger pointing (22.92%).

In our case study, considering all the possible levels for the labeling attributes, there are 2 (number of hands) × 3 (mode of hands) × 6 (direction of motion) × 2 (amplitude) × 3 (height of execution) × 4 (hand shape) = 864 theoretical combinations. Actually, we observed only 52 of these combinations.

3) Gestures clustering

The highest values of IG were found for the attributes 'Direction of motion' and 'Hand shape' (Table I), which led us to cluster the 52 combinations just according to these two attributes.

Moreover, we applied similarity rules by grouping gestures having the palm opened toward the direction of motion with those with the palm opened toward the direction opposite to that of motion.

Adopting this strategy, the original 52 different combinations were finally grouped into 10 gesture clusters (Fig. 4).

The occurrences of gesture clusters for each referent are shown in Table II.

4) Agreement analysis

We obtained the following agreement rate for each referent: *Next, AgR* = .246; *Previous, AgR* = .256; *Go up, AgR* = .542; *Go down, AgR* = .502.

The V_{rd} statistics confirmed that the set of referents is consistent because all the agreement rates are significantly greater than zero (*Next*, $V_{rd}(1) = 1100$, $p = .001$; *Previous*, $V_{rd}(1) = 1144$, $p = .001$; *Go Up*, $V_{rd}(1) = 2418$, $p = .001$; *Go Down*, $V_{rd}(1) = 2242$, $p = .001$).

C. Vocabularies Definition

The elicitation phase returned 10 gesture clusters. The combinations of the gesture clusters for all the referents (see Table II) leads to 5 x 6 x 3 x 4 = 360 possible vocabularies. Applying the "conflict set" constraint, the possible vocabularies reduce to 36 (Fig. 5). For example, the gesture cluster "backward opened palm" (acronym BOP, Fig. 5) has been utilized for both *Next* and *Previous* referent. This gesture cluster is assigned to *Previous* (the gesture clusters removed are highlighted in red, Fig. 5) because for this referent, a higher number of proposals was obtained. Then, applying the "homogeneity constraint," we further reduce the number of vocabularies to five (Fig. 5). For example, combination 1 (Table C, Fig. 5) cannot be accepted as it includes gesture clusters with different attribute "hand shape" and is hence excluded (highlighted in red). For the sake of brevity, in Fig. 5, Table D only the accepted combinations and one (i.e., combination 1) of the unacceptable combinations are shown. Finally, implementing the "reversibility constraint," only three candidate vocabularies remain, namely Vocabulary A, B and C (Table D, Fig. 5). For instance, combination II (Table D, Fig. 5) cannot be accepted because 'forward opened palm' for the referent *Next* is not the opposite gesture of 'left to right opened palm' associated to *Previous*. Whereas, combination I can be accepted (Table D, Fig. 5) because 'right to left opened palm' (RLOP, referent *Next*) is 'opposite' to 'left to right opened palm' (LROP, referent *Previous*) and 'bottom-up opened palm' (BUOP, referent *Go up*) is 'opposite' to 'top-down opened palm' (TDOP, referent *Go down*).

We ranked the three candidate vocabularies resulting from this filtering process in order of guessability *G* (Vocabulary A, *G*=0.49; Vocabulary B, *G* =0.42; Vocabulary C, *G* =0.21). In this case study, the most intuitive vocabulary is Vocabulary A (Fig. 6).

TABLE I

ATTRIBUTE RANKING: THE HIGHER THE INFORMATION GAIN (IG), THE BETTER THE ATTRIBUTE DIFFERENTIATION CAPABILITY. IN THE TABLE ARE HIGHLIGHTED IN GREY THE ATTRIBUTES FOR WHICH THE HIGHEST VALUES OF IG WERE COMPUTED

Attribute	Information gain (IG)
Direction of motion	1.361
Hand shape	0.179
Height of execution	0.068
Mode of hand	0.038
Number of hands	0.003
Amplitude	0.001

Table A. Original combinatorial space of gesture vocabularies

Referent	Number of proposals	Gesture class	Combinations
Next	32	RLOP	5
	26	LROP	
	23	FOP	
	15	FFP	
	2	BOP	
Previous	31	RLOP	6
	31	BFP	
	27	LROP	
	8	BOP	
	3	FOP	
	1	TDOP	
Go up	69	BUOP	3
	20	BUFP	
	11	TDOP	
Go down	67	TDOP	4
	19	TDFP	
	11	BUOP	
	1	BUFP	
Total combinations			5x6x3x4=360

Conflict set constraint

Table B. Combinatorial space after implementing the conflict set constraint

Referent	Number of proposals	Gesture class	Combinations
Next	32	RLOP	3
	26	LROP	
	23	FOP	
	15	FFP	
	2	BOP	
Previous	31	RLOP	3
	31	BFP	
	27	LROP	
	8	BOP	
	3	FOP	
	1	TDOP	
Go up	69	BUOP	2
	20	BUFP	
	11	TDOP	
Go down	67	TDOP	2
	19	TDFP	
	11	BUOP	
	1	BUFP	
Total combinations			3x3x2x2=36

Homogeneity constraint

Table C. Combinatorial space after implementing the conflict set and the homogeneity constraint

Ref.	Comb. 1	Comb. 2	Comb. 3	Comb. 4	Comb. 5	Comb. 6	Comb. …	Comb. 36
Next	RLOP	RLOP	FOP	FFP	RLOP	FOP		
Previous	BFP	LROP	LROP	BFP	BOP	BOP		
Go up	BUOP	BUOP	BUOP	BUFP	BUOP	BUOP		
Go down	TDOP	TDOP	TDOP	TDFP	TDOP	TDOP		
Total remaining combinations								5

Reversibility constraint

Table D. Combinatorial space after implementing the conflict set, the homogeneity, and the reversibility constraint

Referent	Comb. I	Comb. II	Comb. III	Comb. IV	Comb. V
Next	RLOP	FOP	FFP	RLOP	FOP
Previous	LROP	LROP	BFP	BOP	BOP
Go up	BUOP	BUOP	BUFP	BUOP	BUOP
Go down	TDOP	TDOP	TDFP	TDOP	TDOP
Total remaining combinations					3
	VOCABUL. A		*VOCABUL. B*		*VOCABUL. C*

Fig. 5. Filtering process for the identification of the optimal gesture vocabulary. After implementing: conflict set, homogeneity and reversibility constraints, the original combinatorial space of vocabularies decreases from 360 (Table A) to 3 (Table D) possible vocabularies. The acronyms utilized in the figure are described in Fig. 4.

D. Validation procedure

By applying the proposed validation procedure, we checked whether vocabulary A was not only the most intuitive but also satisfied other requirements, such as the fatigue.

We recruited a new users' cohort: twelve participants (average age 29.25 years, SD 3.94 years), all males and right-handed. They were Mechanical Engineering students, nine of them held an MS. The sequence presented 40 queries, i.e., 10

TABLE II:

GESTURE CLUSTERS OCCURRENCES FOR TO EACH REFERENT

Clusters of gestures	*Next*	*Previous*	*Go up*	*Go down*
RLOP	31	31	-	-
LROP	26	27	-	-
FOP	23	3	-	-
FFP	15	-	-	-
BOP	2	8	-	-
BFP	-	31	-	-
TDOP	-	1	11	67
BUOP	-	-	69	11
TDFP	-	-	-	19
BUFP	-	-	20	1

For instance, with reference to the referent "go up", 100 gesture clusters were collected: 69 of these executed the referent "go up" with "bottom-up opened palm", 20 with "bottom-up finger pointing", and 11 with "top-down opened palm".

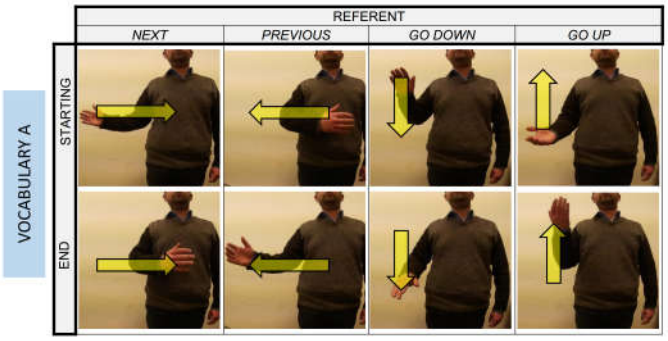

Fig. 6. An example of the gesture vocabulary (A) finally obtained after implementing the conflict set, the homogeneity and the reversibility constraints.

repetitions for each referent, the time interval between each test for a user was one week.

The values of consumed endurance (CE), were positively tested for normality using the Shapiro Wilk test and the assumption of sphericity was met (Mauchly's test: $\chi^2(2)=1.10$, p=0.580). The ANOVA for the within-subject variable showed a significant effect ($F(2,22)=6.355$ p=0.007). Pairwise comparisons showed that Vocabulary A produces a significantly lower CE than Vocabulary B (mean value 31.3±10.8 % vs. 44.3±12.1 %, p=0.016) and Vocabulary C (mean value 45.1±10.7 %, p=0.023).

The "ease of use" requirement was met for all the vocabularies (median value 6 for all). The same result was found for "goodness" (median value 6 for A and C, 6.5 for B). However, statistical analyses revealed that there are no significant differences among the three vocabularies as for "ease of use" (Friedman test, $\chi^2(2)=1.389$, p=0.499) and "goodness" (Kruskal-Wallis test, $\chi^2(2)=4.185$, p=0.123). Also for memorability, there are not statistically significant differences (Kruskal-Wallis test, $\chi^2(2)=0.179$, p=0.914).

Based on these results, we can argue that vocabulary A is not only the most intuitive, but it is preferable also in terms of fatigue, measured through the consumed endurance. Furthermore, none of the other vocabularies has better attributes of memorability, ease of use, and goodness. Then, we can conclude that in our test case, i.e., maintenance digital-manual browsing, vocabulary A should be used.

V. Discussion

A. Framework

The proposed framework, starting from the procedure proposed by Nielsen *et al.* [8], introduces specific elements to minimize the effects of legacy [26] and performance bias [28]. Furthermore, in order to increase the importance of memorability and fatigue, our approach requires a higher number of repetitions for each referent than in previous studies ([26], [27], [45–47], [20]). For the same reason, we reduced the time interval between two consecutive referents in the elicitation step.

This framework can be utilized for user interfaces in different application scenarios. Every scenario provides different user requirements that orientate the choice of the vocabulary during the validation procedure. Within the same scenario, the framework can be used to find different vocabularies for different tasks.

We believe that the framework is scalable up to a reasonable number of commands. The duration of the tests may increase, but it is possible to control it by adjusting the number of repetitions. Moreover, the combinatorial growth of vocabularies can be managed through the constraints applied in the composition phase. For instance, in our case study, the conflict set constraint reduced the number of combinations from 360 to 36, the homogeneity constraint from 36 to 5 and, finally, the reversibility constraint from 5 to 3.

We used the recently introduced CE metric to assess fatigue objectively, thus replacing subjective evaluations such as the Borg CR10 and the NASA-TLX questionnaires [29], [30], and expert's analysis [8]. Overall CE metric was effective to take into account the user's fatigue. However, we found some limitation in CE as the noise in joint detection that required tuning the SkeletonBufferSize parameter to guarantee the correct joint speed assessment. Furthermore, CE cannot take into account other relevant fatigue elements, such as the exertion of hand muscles for the finger movements, the trunk flexion, the neck flexion or extension, and so on.

Given the proposed framework modularity, each step can be easily updated and customized with other algorithms. For instance, we proposed a set of labels in "gesture labeling" (see B.2), but the designer could customize it according to the observed proposals. Moreover, other strategies of fatigue measurements can be easily included in the framework.

Requirements on gesture recognition capability, due to hardware and software technologies, were not explicitly defined in our framework to maintain it device independent.

Although the elicitation approach, definitively cannot conclude what gestures are truly good for users, it allows exploring users' preferences. Furthermore, the proposed framework provides additional quantitative/qualitative indicators that guide the designer to choose the vocabulary thoughtfully for the specific application.

B. Case study

The proposed framework was successfully used in our case study, which allowed us to assess strong and weak points of the framework.

We verified the goodness of the similarity rules used in gesture clustering. Indeed, agreement rates for labeled gestures before applying similarity rules were significantly smaller than those computed for gesture clusters, in accordance with the findings of previous studies [45], [48], [19].

Users preferred reversible gestures to execute referents that have an opposite effect, which is consistent with results obtained by Piumsomboon *et al.* [20], Ooi *et al.* [49] and Silpasuwanchai and Ren [50]. In addition, according to Piumsomboon *et al.* [20], Vocabulary A includes the gestures "swipe left to right" for the referent "*previous*" and "swipe right to left" for the referent "*next*."

The analyses of "ease of use," "goodness of matching," and "memorability" do not evidence any statistically significant differences among the three candidate vocabularies because they are already the result of a strict selection/filtering process and the user is not able to prefer one of them.

The results about CE show that the most intuitive vocabulary (A) is associated with the lowest fatigue. In fact, vocabulary A involves a narrow rotation of the glenohumeral joint that leads

to reduced movements of the center of mass of the kinematic chain of the upper limbs which imply reduced values of the consumed endurance CE. On the contrary, Vocabulary B and C require, for some referents, lifting both the forearm (as for vocabulary A) and the arm.

In conclusion, this work proposes a general framework to design mid-air gesture-based interfaces. It was successfully applied in the case study of maintenance digital-manual browsing. The systematic and wide breadth approach adopted in the study makes the proposed design methodology easily extendable to other contexts where mid-air gestures might turn useful. Furthermore, this framework can also be applied, with just little adjustments, to other types of vocabularies, not only to mid-air gestures. For instance, it was successfully used for the definition of an optimal vocabulary of graphical symbols to be used in Augmented Reality to represent maintenance instructions [51]. Finally, it is worthy to note that the phases A and D can also be used as a stand-alone procedure for a comparative evaluation of existing vocabularies in gesture interfaces, without deriving them from a user elicitation.

REFERENCES

[1] R. Aigner, D. Wigdor, H. Benko, M. Haller, D. Lindbauer, A. Ion, S. Zhao, and J. Koh, "Understanding mid-air hand gestures: A study of human preferences in usage of gesture types for hci," *Microsoft Research TechReport MSR-TR-2012-111*, vol. 2, 2012.

[2] A. T. Cabreira and F. Hwang, "An analysis of mid-air gestures used across three platforms," in *Proceedings of the 2015 British HCI Conference*, 2015, pp. 257–258.

[3] T. Baudel and M. Beaudouin-Lafon, "Charade: remote control of objects using free-hand gestures," *Communications of the ACM*, vol. 36, no. 7, pp. 28–35, 1993.

[4] H. Witt, T. Nicolai, and H. Kenn, "Designing a Wearable User Interface for Hands-free Interaction in Maintenance Applications," in *Proceedings of the 4th Annual IEEE International Conference on Pervasive Computing and Communications Workshops*, 2006, pp. 652–655.

[5] B. Schwald and B. De Laval, "An augmented reality system for training and assistance to maintenance in the industrial context," *Journal of WSCG*, vol. 11, no. 1–3, 2003.

[6] M. Fiorentino, R. Radkowski, C. Stritzke, A. E. Uva, and G. Monno, "Design review of CAD assemblies using bimanual natural interface," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 7, no. 4, pp. 249–260, 2013.

[7] M. Fiorentino, A. E. Uva, G. Monno, and R. Radkowski, "Natural interaction for online documentation in industrial maintenance," *International Journal of Computer Aided Engineering and Technology*, vol. 8, no. 1–2, pp. 56–79, 2016.

[8] M. Nielsen, M. Störring, T. Moeslund, and E. Granum, "A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI," in *Gesture-Based Communication in Human-Computer Interaction*, vol. 2915, A. Camurri and G. Volpe, Eds. Springer Berlin Heidelberg, 2004, pp. 409–420.

[9] W. K. English, D. C. Engelbart, and M. L. Berman, "Display-selection techniques for text manipulation," *IEEE Transactions on Human Factors in Electronics*, no. 1, pp. 5–15, 1967.

[10] Microsoft, "Human Interface Guidelines v2.0, http://download.microsoft.com/download/6/7/6/676611B4-1982-47A4-A42E-4CF84E1095A8/KinectHIG.2.0.pdf." 2014.

[11] M. Fiorentino, A. E. Uva, M. Gattullo, S. Debernardis, and G. Monno, "Augmented reality on large screen for interactive maintenance instructions," *Computers in Industry*, vol. 65, no. 2, pp. 270–278, 2014.

[12] A. E. Uva, M. Gattullo, V. M. Manghisi, D. Spagnulo, G. L. Cascella, and M. Fiorentino, "Evaluating the effectiveness of spatial augmented reality in smart manufacturing: a solution for manual working stations," *The International Journal of Advanced Manufacturing Technology*, vol. 94, no. 1–4, pp. 509–521, 2018.

[13] H. I. Stern, J. P. Wachs, and Y. Edan, "Optimal consensus intuitive hand gesture vocabulary design," in *Semantic Computing, 2008 IEEE International Conference on*, 2008, pp. 96–103.

[14] T. Kirishima, K. Sato, and K. Chihara, "Real-time gesture recognition by learning and selective control of visual interest points," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 351–364, 2005.

[15] K. H. Munk, "Development of a gesture plug-in for natural dialogue interfaces," in *International Gesture Workshop*, 2001, pp. 47–58.

[16] K. Kahol, K. Tripathi, and S. Panchanathan, "Documenting motion sequences with a personalized annotation system," *IEEE Multimedia*, vol. 13, no. 1, pp. 37–45, 2006.

[17] V. M. Manghisi, M. Fiorentino, M. Gattullo, A. Boccaccio, V. Bevilacqua, G. L. Cascella, M. Dassisti, and A. E. Uva, "Experiencing the Sights, Smells, Sounds, and Climate of Southern Italy in VR," *IEEE computer graphics and applications*, no. 6, pp. 19–25, 2017.

[18] V. M. Manghisi, A. E. Uva, M. Fiorentino, M. Gattullo, A. Boccaccio, and G. Monno, "Enhancing user engagement through the user centric design of a mid-air gesture-based interface for the navigation of virtual-tours in cultural heritage expositions," *Journal of Cultural Heritage*, vol. 32, pp. 186–197, 2018.

[19] M. R. Morris, J. O. Wobbrock, and A. D. Wilson, "Understanding users' preferences for surface gestures," in *Proceedings of graphics interface 2010*, 2010, pp. 261–268.

[20] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn, "User-Defined Gestures for Augmented Reality," in *Human-Computer Interaction – INTERACT 2013*, vol. 8118, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds. Springer Berlin Heidelberg, 2013, pp. 282–299.

[21] J. O. Wobbrock, H. H. Aung, B. Rothrock, and B. A. Myers, "Maximizing the Guessability of Symbolic Input," in *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 2005, pp. 1869–1872.

[22] H. I. Stern, J. P. Wachs, and Y. Edan, "Designing hand gesture vocabularies for natural interaction by combining psycho-physiological and recognition factors," *International Journal of Semantic Computing*, vol. 2, no. 01, pp. 137–160, 2008.

[23] H. I. Stern, J. P. Wachs, and Y. Edan, "Human factors for design of hand gesture human-machine interaction," in *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, 2006, vol. 5, pp. 4052–4056.

[24] J. F. Hessam, M. Zancanaro, M. Kavakli, and M. Billinghurst, "Towards optimization of mid-air gestures for in-vehicle interactions," in *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, 2017, pp. 126–134.

[25] A. Pereira, J. P. Wachs, K. Park, and D. Rempel, "A user-developed 3-D hand gesture set for human–computer interaction," *Human factors*, vol. 57, no. 4, pp. 607–621, 2015.

[26] M. R. Morris, A. Danielescu, S. Drucker, D. Fisher, B. Lee, m. c. schraefel, and J. O. Wobbrock, "Reducing Legacy Bias in Gesture Elicitation Studies," *interactions*, vol. 21, no. 3, pp. 40–45, May 2014.

[27] J. O. Wobbrock, M. R. Morris, and A. D. Wilson, "User-defined Gestures for Surface Computing," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 1083–1092.

[28] J. Ruiz and D. Vogel, "Soft-Constraints to Reduce Legacy and Performance Bias to Elicit Whole-body Gestures with Low Arm Fatigue," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 3347–3350.

[29] G. Borg, *Borg's perceived exertion and pain scales.* Champaign, IL, US: Human kinetics, 1998.

[30] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Human mental workload*, vol. 1, no. 3, pp. 139–183, 1988.

[31] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2006, vol. 50, no. 9, pp. 904–908.

[32] B. R. Jensen, B. Laursen, and G. Sjøgaard, "Aspects of shoulder function in relation to exposure demands and fatigue–a mini review," *Clinical Biomechanics*, vol. 15, pp. S17–S20, 2000.

[33] L. Chittaro and R. Sioni, "An electromyographic study of a laser pointer-style device vs. mouse and keyboard in an object arrangement task on a large screen," *International Journal of Human-Computer Studies*, vol. 70, no. 3, pp. 234–255, 2012.

[34] V. M. Manghisi, A. E. Uva, M. Fiorentino, V. Bevilacqua, G. F. Trotta, and G. Monno, "Real time RULA assessment using Kinect v2 sensor," *Applied ergonomics*, vol. 65, pp. 481–491, 2017.

[35] L. McAtamney and E. Nigel Corlett, "RULA: a survey method for the investigation of work-related upper limb disorders," *Applied ergonomics*, vol. 24, no. 2, pp. 91–99, 1993.

[36] J. D. Hincapié-Ramos, X. Guo, P. Moghadasian, and P. Irani, "Consumed Endurance: A Metric to Quantify Arm Fatigue of Mid-air Interactions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 1063–1072.

[37] W. Rohmert, "Ermittlung von Erholungspausen für statische Arbeit des Menschen," *European Journal of Applied Physiology and Occupational Physiology*, vol. 18, no. 2, pp. 123–164, 1960.

[38] ACGIH, "Upper Limb Localized Fatigue: TLV(R) Physical Agents 7th Edition Documentation." Report number 7DOC-782; ACGIH; Cincinnati, Ohio, 2016.

[39] J. Blake, *Natural User Interfaces in .Net*. Manning Publications, 2012.

[40] M. Frisch, J. Heydekorn, and R. Dachselt, "Investigating multi-touch and pen gestures for diagram editing on interactive surfaces," in *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, 2009, pp. 149–156.

[41] V. R. Preedy, Ed., *Handbook of Anthropometry*. Springer US, 2012.

[42] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[43] R.-D. Vatavu and J. O. Wobbrock, "Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 1325–1334.

[44] W. et al., "AGreement Analysis Toolkit , http://depts.washington.edu/aimgroup/proj/dollar/agate.html." 2016.

[45] H. Wu and J. Wang, "User-Defined Body Gestures for TV-based Applications," in *Digital Home (ICDH), 2012 Fourth International Conference on*, 2012, pp. 415–420.

[46] W. Fikkert, P. van der Vet, G. van der Veer, and A. Nijholt, "Gestures for Large Display Control," in *Gesture in Embodied Communication and Human-Computer Interaction*, vol. 5934, S. Kopp and I. Wachsmuth, Eds. Springer Berlin Heidelberg, 2010, pp. 245–256.

[47] R.-D. Vatavu, "User-defined Gestures for Free-hand TV Control," in *Proceedings of the 10th European Conference on Interactive Tv and Video*, 2012, pp. 45–48.

[48] G. Bailly, T. Pietrzak, J. Deber, and D. J. Wigdor, "Métamorphe: augmenting hotkey usage with actuated keys," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 563–572.

[49] B. Ooi, C. Wong, I. Tan, and C. Lee, "Towards Natural Gestures for Presentation Control Using Microsoft Kinect," in *Advances in Multimedia Information Processing – PCM 2014*, vol. 8879, W. Ooi, C. M. Snoek, H. Tan, C.-K. Ho, B. Huet, and C.-W. Ngo, Eds. Springer International Publishing, 2014, pp. 258–261.

[50] C. Silpasuwanchai and X. Ren, "Designing concurrent full-body gestures for intense gameplay," *International Journal of Human-Computer Studies*, vol. 80, pp. 1–13, 2015.

[51] G. W. Scurati, M. Gattullo, M. Fiorentino, F. Ferrise, M. Bordegoni, and A. E. Uva, "Converting maintenance actions into standard symbols for Augmented Reality applications in Industry 4.0," *Computers in Industry*, vol. 98, pp. 68–79, 2018.

Footnotes:

Manuscript received

Authors are with the Department of Mechanics, Mathematics and Management, Polytechnic Institute of Bari, Italy. (e-mail {antonio.uva, michele.fiorentino, vitomodesto.manghisi, antonio.boccaccio, saverio.debernardis, michele.gattullo, giuseppe.monno}@poliba.it)

List of figure captions:

Fig. 1. Flow-chart of the proposed framework.
Fig. 2. The six general attributes utilized to label/classify (Step B.2) the collected gesture proposals and successively implemented in the case study.
Fig. 3. Experimenter view of VocVal. Participants can see only the required command.
Fig. 4. Gesture clusters obtained combining the 2 most significant attributes and applying the similarity rules.
Fig. 5. Filtering process for the identification of the optimal gesture vocabulary. After implementing: conflict set, homogeneity and reversibility constraints, the original combinatorial space of vocabularies decreases from 360 (Table A) to 3 (Table D) possible vocabularies. The acronyms utilized in the figure are described in Fig. 4.
Fig. 6. An example of the gesture vocabulary (A) finally obtained after implementing the conflict set, the homogeneity and the reversibility constraints.