UCLouvain

LOUVAIN
School of Management

Louvain School of Management

The managerial implications behind camera-acquired 3D gestures classifications to control a computer system.

Author: Gil Decottignies
Supervisor: Jean Vanderdonckt
Academic year 2021-2022
Dissertation for the Master of Business Engineering
Daytime schedule

Acknowledgements

First of all, I would like to thank my supervisor, Professor Jean Vanderdonckt, for his help, his availability, and his kindness.

Secondly, I would like to thank Santiago Villareal-Narvaez, for proofreading this thesis, for his help and his precious advice.

Finally, I would like to thank my mother Laurence, my sister Audrey, my best friend Clotilde, and my roommate Natale, for the unfailing support they gave me in the writing of this thesis.

Table of contents

1. INTRODUCTION 1
1.1 Contextualisation 2
1.2 Motivation 3
1.3 Research Question 5
2. RELATED WORK 6
2.1 Human-Computer Interaction Interfaces requirements 6
2.1.1 Gesture Elicitation Study (GES) 6
2.1.1.1 The guessability method 7
2.1.1.2 The Choice-based Elicitation Method 11
2.1.1.3 The "Intuitive and Ergonomic" Method 12
2.1.1.4 The Wizard-of-Oz technique 12
2.1.1.5 Gesture criteria 13
2.1.2 Interface Functionalities and Performance 13
2.1.2.1 General Basic Requirements 14
2.1.2.2 General Individual Level Computing Domains Requirements 16
2.1.2.3 General Collaborative Level Computing Domains Requirements 21
2.2 Existing gesture classifications 23
2.2.1 Efron 24
2.2.2 Kendon 24
2.2.3 McNeill 24
2.2.4 Quek et al. 25
2.2.5 Karam & Schraefel 26
2.2.6 Aigner et al. 29
2.2.7 Vafaei 31
2.2.7.1 Gesture mapping 32
2.2.7.2 Physical characteristics 33
2.2.8 Hiyadi 36
2.3 Gesture information representation 37
2.3.1 McAweeney et al. 37
2.3.2 Bernsen and Dybkjær 39
3. GES REVIEW IN CHOSEN APPLICATION DOMAINS 44
3.1 Gesture set selection 44
3.2 Taxonomy used to classify the gesture set 48
3.2.1 Classification hypotheses 50

3.3 Gesture set analysis 51
3.3.1 Nature 51
3.3.2 Form 52
3.3.3 Binding 53
3.3.4 Temporal 53
3.3.5 Context 54
3.3.6 Complexity 54
3.3.7 Body part 55
3.3.8 Handedness 56
3.3.9 Handshape 56
3.4 Main issues when classifying gestures 57
3.5 A taxonomy for classifying 3D-acquired gestures 58
4. MANAGERIAL IMPLICATIONS 60
4.1 Exploratory survey analysis 60
4.2 Challenges behind interfaces requirements 63
4.2.1 Price 63
4.2.2 Responsiveness 64
4.2.3 User adaptability and feedback 64
4.2.4 Learnability 64
4.2.5 Accuracy 64
4.2.6 Low mental load 65
4.2.7 Intuitiveness 65
4.2.8 Comfort 65
4.2.9 Lexicon size and multi-hand systems 65
4.2.10 "Come as you are" 65
4.2.11 Reconfigurability 66
4.2.12 Gesture spotting and immersion syndrome 66
4.3 Computing domains requirements 66
4.4 Gesture representation in "quick learning guide" 67
4.5 Implications when classifying gestures post GES 68
5. LIMITS 70
6. CONCLUSION 71
7. BIBLIOGRAPHY 72
8. APPENDIX 80

1. INTRODUCTION

This master thesis is composed of five main parts: Introduction, Related Work; GES review; Managerial implications and Conclusion and Limitations.

The purpose of the introduction is to provide some background on the topic; to explain my motivations for proposing a taxonomy that might be useful to managers when they wish to classify their own actions, as well as to provide the managerial implications behind it; and to describe my research question.

Then, the related work section describes both the requirements related to the development of communication interfaces in general, the different possibilities to carry out a gesture elicitation study, as well as a kind of chronological summary of the evolution of the different possible classifications. In addition, the representation of post-study gestures will also be evaluated.

Afterwards, the review of 42 studies describes the most common gestures across the various application areas chosen, explains how these gestures were classified in order to be analysed according to all the selected criteria, and finally proposes a taxonomy that can be used by managers to classify the gestures acquired by camera when they conduct their own studies.

Thereupon, the managerial implications contain the results of an exploratory survey I conducted as part of this dissertation, the different challenges behind the needs when developing a new system, as well as the managerial implications that arise from the proposed taxonomy.

Finally, the last part summarises both the wider managerial implications behind the classification of 3D camera-acquired gestures and the development of this type of interface, as well as reviewing the limitations of this thesis.

1.1 Contextualisation

In an increasingly connected world, human-computer interactions (HCI) are constantly rising given that technological innovations continue to emerge and revolutionise our lives. Whether it be game consoles, tablets, computers, phones or TVs, products are indeed constantly evolving with the sole aim of offering more comfort and functionalities to the user. However, communicating and using all these new machines is not always an easy task for humans as they have to learn and adapt to using these machines which can take long periods of time.

To remedy these potential problems of man-machine interaction, humans have developed communication interfaces (camera, smart glasses, smart ring, ...) allowing to detect, analyse and use the gestures made by the users in order to carry out various actions and control those machines. These interactive technologies are therefore there to facilitate communication between man and the machine he uses, technologies among which the best known are certainly the Kinect developed by Microsoft, the Google Glass or even Leap Motion. As a matter of fact, gestures (body movement, facial expression and finger pointing) are probably the most intuitive and natural type of communication as we use them since the earliest stages of our development. Moreover, these new technologies that help HCI are proven to enhance the power, speed, efficiency and realism of the computer systems that use them (Wachs et al., 2011, p. 60).

The fields of application in which we find HCI are numerous and varied, among which we find Augmented and Virtual Reality (one of the largest area), Desktop and Tablet PC Applications, Three Dimensional Displays, Ubiquitous Computing and Smart Environments, Games, Pervasive and Mobile Interfaces, Telematics, Adaptive Technology, Communication Interfaces and finally Gesture Toolkits (Karam & Schraefel, 2005, pp. 13-20). Furthermore, Wachs et al. (2011, pp. 60, 62) explain that hand-gesture interfaces are useful especially in four domains of application thanks to the main advantages they provide. The first domain is medical systems and assistive technologies because touchless interfaces enable access to information while maintaining total sterility. The second one is crisis management and disaster relief as it facilitates the exploration of big data thanks to 3D interaction. The third domain is entertainment (including Games) because these interfaces help overcoming physical handicaps. The last one is human-robot interaction with the main motivation being that communication seems as much as possible as natural human dialogue. In addition to these domains, Yong (2014, pp. 65-71) described in his thesis several Individual and Collaborative Level General Computing Domains that are Productivity, Media Consumption, Communication and

Augmentation which will be explained in detail later in this work. Moreover, more recently, Koutsabasis and Vogiatzidakis (2019, pp. 1749-1753) added new areas such as Text Entry, Cultural Heritage, Therapeutic Assistance and TVs; and confirmed that others were still relevant such as Ubiquitous and Smart Environment, Sharing among devices (big data), 3D objects and Space, Operating Rooms, Accessibility (by gamification) and Distant Displays more globally.

1.2 Motivation

Having chosen to specialise in innovation management during my master's degree in management engineering, it seemed quite appropriate to choose a thesis topic related to an innovative technology. This is why, when choosing the subject, I was interested in this topic about Human-Computer Interaction. Since I was a kid, I always liked to be in contact with technology, computers, smartphones, game consoles, etc. and I remember one day seeing a game on TV that caught my attention as it allowed you to control your character in an open world simply by using a wireless controller in your hand. The equivalent of the Nintendo Wii that I experienced when it came out a few years after seeing this game on TV and which was my first real experience of gestural communication with a machine. Then I was able to try out the Microsoft Kinect at a friend's house to play Just Dance, and then I had the joy of testing myself in virtual reality, but still in the entertainment and gaming field. I was therefore very curious to discover to what extent this kind of human-computer interface can help us in today's society and this motivated me to do my research in order to discover the various fields of application that I could analyse. However, I did not intend to neglect the field of entertainment, which is the one I identify with the most.

In order to be as complete and coherent as possible, I cannot deal with all the above-mentioned areas of application in this thesis. Therefore, I have chosen to focus mainly on the main application areas proposed by Wachs et al. (2011, p. 60, 62) in their study, and confirmed by Koutsabasis and Vogiatzidakis (2019, pp. 1749-1753), namely **Medical Systems and Assistive Technologies** (or **Operating Rooms**), **Crisis Management and Disaster Relief** (or **Sharing Among Devices**), **Entertainment** (or **Accessibility**) and **Human-Robot Interaction** as well as **TVs** (Koutsabasis & Vogiatzidakis, 2019, p. 1752). Besides, I also take into consideration in the requirement’s part the Computing Domains mentioned by Yong in his thesis (2014, p. 65-69), which are as a reminder **Productivity, Media Consumption, Communication** and **Augmentation**.

The focus of this thesis is on cameras that serve as communication interfaces between man and computer because they do not require the use of any wearable device (ring, glasses, lenses or other) - although both can be combined - and they detect gestures in a three-dimensional (3D) environment. In addition, cameras have the advantage of being able to detect the entire range of body movements, from head to toe, and not necessarily focus on a specific part of the body. Those two characteristics of cameras are important as they ensure access to information while maintaining total sterility, ease the exploration of big data and also help to overcome physical handicaps. Consequently, camera interfaces seem perfect to provide the advantages that make easy conventional human-machine interaction.

To better understand how people would use such camera-based communication interfaces, gesture elicitation studies were conducted to find out which gestures were repeated most often by users to perform the same task. These studies also took into account several characteristics and notably the ease of performing and retaining the gesture to perform a specific action. Next to that, several taxonomies were created and used to classify the collected gestures in these elicitation gesture studies. Each of these classifications has its advantages and disadvantages, and the aim of this thesis will be to give a sort of account of these advantages and disadvantages so that a manager wishing to use one of these interfaces or to develop a new one can easily understand what is important to consider or not in the detection of movements.

Furthermore, as this thesis is part of my master's degree in business engineering, it will focus on the use of cameras to control an information system. Although there is still no universal definition of what an information system is, one of the most complete I found is provided by Vladimir Zwass in the Encyclopaedia Britannica (2022):

Information system, an integrated set of components for collecting, storing, and processing data and for providing information, knowledge, and digital products. Business firms and other organizations rely on information systems to carry out and manage their operations, interact with their customers and suppliers, and compete in the marketplace. Information systems are used to run interorganizational supply chains and electronic markets. For instance, corporations use information systems to process financial accounts, to manage their human resources, and to reach their potential customers with online promotions. (Zwass, 2020, November 2)

According to Legrenzi's research (2016), who tried to provide a definition coming from all the various definitions that currently exist, a simpler definition would be "*Le système d'information représente l'ensemble des ressources internes ou externes – utilisateurs, outils, données – qui contribuent au traitement (numérique ou non) de l'information*"[1].

1.3 Research Question

Given the above context, the research question that guided my work is: ***How to classify the gestures that can be acquired by camera to control a computer system? What are the managerial implications behind these classifications?***

In this one, we find the information system aspect as well as managerial implications linked to the classification of gestures acquired by camera. Nevertheless, to avoid the double question, I reworked the topic to obtain the following theme: ***The managerial implications behind camera-acquired 3D gestures classifications to control a computer system***.

Therefore, in the course of this thesis I will have to make a selection of already existing gesture classifications in order to summarise their benefits and drawbacks, select a few camera elicitation studies according to the pre-chosen application areas and classify the collected gestures according to the chosen classifications.

If we put all this into the context of the COVID pandemic that we have just gone through, it seems relevant to ask why these kinds of communication interfaces were not used more during the pandemic in supermarkets, malls, hospitals, etc. in order to fight against the spread of the virus. Indeed, as these technologies allow remote control of devices without the need to touch screens, touchless technologies that control an electronic device could have limited the rate of contamination in some way.

[1] "*The information system represents all the internal or external resources - users, tools, data - that contribute to the processing (digital or not) of information*". [My translation]

2. RELATED WORK

In this section, I report on previous research in the field of HCI. First, I list the requirements for the development of a new human-machine communication interface according to the previously chosen application areas. Then, I discuss the different classifications that exist to categorise the gestures captured by the cameras as well as to represent acquired gestures.

2.1 Human-Computer Interaction Interfaces requirements

First and foremost, it is important when developing a new technology in HCI to consider the user requirements. As I choose to focus on cameras as communication interfaces between man and machine, the requirements are based notably on mid-air gesture interactions. De facto, mid-air gesture interaction are defined by Vogiatzidakis and Koutsabasis (2018, p. 1) as "*touchless manipulations of digital content or remote devices, based on tracking of body movements, postures and gestures with non-intrusive sensors (or minimally-intrusive, mainly based on computer vision)*". As of today, according to Koutsabasis and Vogiatzidakis (2018, p.1, and 2019, pp. 1749-1753) in mid-air interactions, whether it is with remote devices or digital content, there are neither established design patterns nor universal gesture vocabularies. On the contrary, it is acknowledged that identifying appropriate gestures depends on the context of use and is thus a very important design issue. To do so, user-requirements research is conducted, particularly in the form of gesture elicitation studies, to select the most accepted gestures for target users.

2.1.1 Gesture Elicitation Study (GES)

To start, it seems important to explain several terms that are used in GES. As a matter of fact, there are several terms that must be explained in order to better understand how these studies are designed. Therefore, I will explain the following terms mainly based on Villareal-Narvaez et al. (2020, pp. 856-857):

- **Elicitation** = act of conducting a study by asking target end-users to use gestures to perform referents.

- **Participant** = Representative of the target end-user population that accepts to participate voluntarily to the GES.
- **Referent** = Expected result of an action that is presented to participants (Koutsabasis and Vogiatzidakis, 2019, p. 1754). For instance, it could be "Turn On/Off TV", "Turn On/Off Light", etc.. Can also be called Function according to Villareal-Narvaez et al. (2020, p 857), who describe a Referent/**Function** as *"A feature of the user interface that can be controlled independently using a command".* In short, it corresponds to a description of what is going to happen when performing a specific pre-defined gesture.
- **Gesture** = Proposed by the participant, it can be a movement of a part or of the whole body and is a response to a referent.
- **Command =** Signal pre-defined that triggers the execution of a function. In our case, as the signal is a gesture input, thus it is referred to as a "gesture command". For instance, to take the same referent example as before, in response to the "Turn On TV" referent, a participant may propose a snap of the fingers.
- **Agreement** = Corresponds to the case in which two or more participants proposed gestures that are considered as equivalent according to a predefined rules package from the target application's point of view.

2.1.1.1 The guessability method

Now that the lexical field of GES is established, let us analyse how such studies are conducted. Koutsabasis and Vogiatzidakis (2019, p. 1754) explain in their work that there are six main steps when conducting a gesture elicitation study "*in line with the guessability method proposed by Wobbrock, Aung, Rothrock, & Myers (2005)*[2]", a method whose purpose is to evaluate and maximise the guessability of symbolic user input:

- ➔ **Step 1**: A list of referents is established by the researcher (s) who presents referents one by one to the participant.
- ➔ **Step 2**: The participant, depending on what has decided researcher(s), suggests one (or several) gesture(s) that fits the referent. Sometimes, researcher(s) can ask the participant to propose a gesture without thinking too much to keep intuitiveness, ease of use, etc.

[2] Wobbrock, J. O., Aung, H. H., Rothrock, B., & Myers, B. A. (2005). Maximizing the guessability of symbolic input. New York, NY: ACM. Retrieved from http://portal.acm.org/citation.cfm?doid=1056808. 1057043

- ➔ **Step 3**: A camera records a video of the gesture proposed by the participant, and if they wish, the researcher can also take notes of their observations simultaneously.
- ➔ **Step 4**: When the participant has finished proposing gestures for the referent, the researcher presents the next one.
- ➔ **Step 5**: When all the referents have been presented, the participants are asked to fill in a questionnaire or to participate in a short summary discussion to pre-evaluate the proposed gestures according to various aspects (memorability, intuitiveness, etc.).
- ➔ **Step 6**: After analysing the data, the researcher establishes the set of gestures.

Besides, in the "LINFO 1311 Human-Computer Interfaces" course, taught by my promoter Professor Jean Vanderdonckt (2021) at UCLouvain, a more detailed explanation on how to perform a GES following this guessability method is given, mainly based on Wobbrock et al.'s (2009) [3]research. This is done in seven steps:

- ➔ **Step 1**: **Prepare set up C= (U, P, E)**
  - ◆ U for User
    - Identify target end-users and create a sampling with those users. It can be a Random Sampling if the whole population is available; a Systematic Sampling if a sufficient number of the population is available, in this case select every X-th member of the population; or a Stratified Sampling if you want to create subgroups according to different criteria (age, health status, gender, …).
    - Define a schedule to experiment with each participant individually.
    - Create a list of different referents for each task and randomise the order of this list for each participant.
  - ◆ P for Platform
    - Fix hardware and software.
    - Ensures their stability throughout the experiment.
    - As gestures must be recorded preferably in LogFiles, identify the right Software to do so.

[3] Wobbrock, J. O., Morris, M. R., & Wilson, A. D. (2009). User-defined gestures for surface computing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1083-1092. https://doi.org/10.1145/1518701.1518866

- E for Environment (physical and psychological)
  - Depending on the device, determine the perfect physical location (lab, specific room, office, etc.).
  - Be sure to be able to stay in the same conditions during the whole experiment with each user.

➔ **Step 2**: **Pre-test phase**

- For the researcher:
  - Choose a differentiation sign (e.g. an ID number) for each participant that you can put on the consent form.
  - Make sure each participant correctly fills-in the documents.
- For each participant:
  - Fill-in a consent file; as well as a demographic form to know if they are used to using the target device, and what their general system experience with this type of device is.
  - Check their motor skills with a test: ask them to join their thumb to each of their other fingers (on the same hand) two times in a row without making a mistake.
  - Present them the device they will have to experiment with. Use a video, a quick start guide, etc.

➔ **Step 3**: **Test phase**

- Present each participant with the random list of referents that you have prepared for them.
- Verify they understand what the task is.
- Present the referent one by one and ask for a suitable gesture.
  - Ask the participant to provide a rating (from 1 to 5 for example) to represent the extent to which they find the proposed gesture to be consistent with the referent.
  - Depending on what you want to do, you can also ask the participant to note other aspects such as memorability, complexity, etc.
- Monitor the time the participant takes to propose the gesture (thinking-time).
- Record a video of the experiment to go further in your analysis.

➔ **Step 4: Post-test phase**

- Ask the participant to complete the IBM Post Study System Usability Questionnaire.
- Have a wrap-up discussion with the participant if you want them to give more information about the experiment (what they like and dislike with the device, etc.). You can also use an open questions questionnaire.

➔ **Step 5**: **Gesture classification and Computation of the agreement among gestures**

- **Gesture Classification**
  - Name the collected gestures in a structured and consistent way: (action verb) + (action) + (mean).
  - Categorise each gesture with a specific ID (e.g., regroup all the swipe right movements)
  - Choose to use one of the existing classifications or create your own classification based on what exists. (Classifications will be discussed in more detail later in this work (see section 2.2 Existing gesture classifications)).
- **Compute the agreement among gestures**
  - Agreement Rate (AR) must be computed for all referents. AR corresponds to the "*number of pairs of participants in agreement with each other divided by the total number of pairs of participants that could be in agreemen*t" (Vanderdonckt, 2021).
  - Co-agreement rate can also be computed for pairs, groups (e.g., male opposed to female) and categories of referents (e.g., basic opposed to advanced).

➔ **Step 6**: **Analysing agreement**

- Sort the referents in descending order of their AR, with error bars indicating the confidence interval (95%) and the gesture category.
- Example:

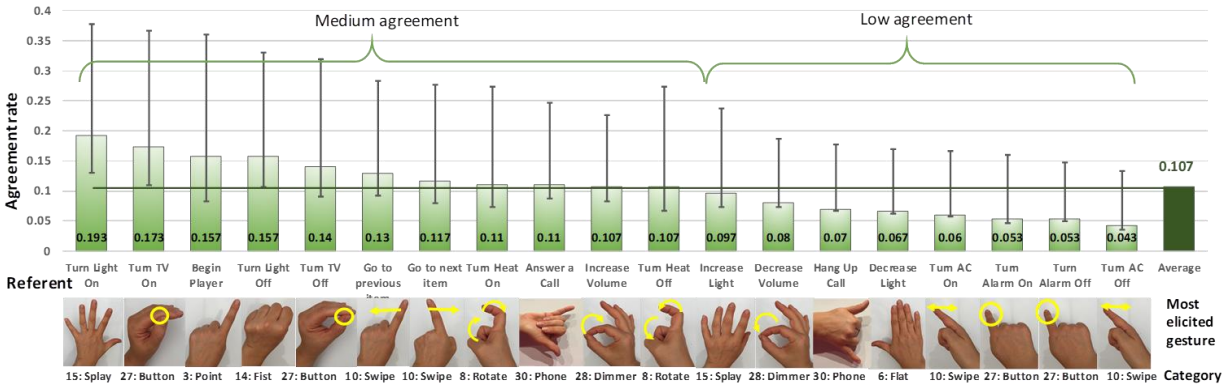


Figure 1-Agreement Rate Analysis Example (Magrofuoco et al., 2019, p. 176987)

➔ **Step 7: Analyse gestures**

- Make a detailed analysis of the gestures acquired during the study according to criteria, agreement rate, observations, wrap-up discussion, questionnaires filled-in by participants, etc.
- Check statistically if there are correlations between the data (ANOVA, …).
- Find a Consensus gesture set based on results.

However, the guessability method is not the only pertinent method to make a GES successful. Vogiatzidakis and Koutsabasis (2018, pp. 6-7) indeed suggest two other methods for conducting a GES.

2.1.1.2 The Choice-based Elicitation Method

The first method is called the choice-based elicitation method. In fact, this method is complementary to the guessability method and has two phases. The first one consists of a classic guessability study as explained above. The second one, which is kind of a refinement technique, only applies to the referents that had a low agreement rate. To re-examine these low agreement referents, users are asked to participate in a new survey where they have to choose the gesture that best fits the task at hand from a predefined gesture list created by the researchers or the users themselves.

2.1.1.3 The "Intuitive and Ergonomic" Method

The second alternative method is the "Intuitive and Ergonomic" Method proposed by Nielsen et al. (2003)[4] quoted by Vogiatzidakis and Koutsabasis (2018, p.7). This type of study aims to design gestures based upon usability principles such as intuitiveness, ease of performing and remembering, ergonomic and metaphorical logic. This method implies to examine beforehand what is the most appropriate interaction technique to be developed regarding the system, as they recognise that mid-air gesture interaction is not the ideal solution for all applications. After doing that, they suggest using scenarios where participants have direct interactions with the researcher or another user by using the most appropriate gestures for the specific function; the idea is to use a human-to-human non-verbal communication approach to generate intuitive gestures for a system. Nevertheless, sometimes when it is needed, the design of the interface or the feedback can be tested using the human-to-computer approach called the Wizard-of-Oz technique. Then, when the collection of gestures and the evaluation of their ergonomic characteristics have been completed, the researcher can benchmark the resulting gesture vocabulary regarding criteria such as guessability, memorability and stress.

2.1.1.4 The Wizard-of-Oz technique

Given that this technique is used in some of the studies I use in the practical part of this thesis and that in some cases researchers have no choice but to use it to develop their system; it seemed to me very relevant to explain more widely what this method actually consists of. The Wizard-of-Oz (WoZ) technique is useful when you want to develop a prototype and is commonly used in HCI research (Dahlbäck et al., 1993[5], quoted by Schiavo et al., 2016, p.93). As do other methods, this technique aims to involve the end-user in the design of systems that are not yet developed (Schiavo et al., 2016, p.93). As reported by Green and Wei-Haas (1985)[6] quoted by Schiavo et al. (2016, p. 94), the first application of the WoZ technique was in the field of natural language processing. Then, this technique has spread to other areas such as HCI, Human-Robot Interaction, Usability, and User-Centred design. The WoZ method enables

[4] Nielsen, M., Moeslund, T.B., Störring, M., & Granum, E. (2003). A procedure for developing intuitive and ergonomic gesture interfaces for man-machine interaction.
[5] Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993): Wizard of Oz Studies: Why and How. In Proc. of the 1st Int. Conference on Intelligent User Interfaces, 193–200.
[6] Green, P., and Wei-Haas, L. (1985): The Wizard of Oz: a tool for rapid development of user interfaces. Final report (Technical report). University of Michigan.

users to operate with a seemingly fully operational system, the missing functions being added by a human operator called “wizard”. The wizard, which is in most of the cases hidden to the users, has the ability to feign the effect of one or more not yet implemented functions from the system. For instance, in the case of a system based on an automatic speech recognizer, the user considers the system fully operational because the system manually responds appropriately thanks to the wizard who listens to the commands announced by the user. The real aim of WoZ prototypes is to enable users to try an external representation of a final system before its final development. The involvement of users can happen at different stages of the design process, whether it be for obtaining their opinions on the Woz prototype or for involving them in features discussion about a future technology (Schiavo et al., 2016, p.94).

2.1.1.5 Gesture criteria

Several times before, pre-defined gesture criteria were mentioned but not really defined. However, it is important to clearly understand what these criteria mean. As a reminder, I often talked about memorability, ease of use, complexity, etc. which are criteria that may be necessary to consider depending on the type of information system you wish to develop.

In their work, Koutsabasis and Vogiatzidakis (2019, p. 1755) explain what are all the dimensions that have been considered in the GES they have used. The three main criteria that come up the most in these studies are gestures that are **easy to perform**, **intuitiveness or natural** (few or no instruction needed) and **memorability.** Next to these three criteria, other dimensions can be important to consider. As a matter of fact, some researchers take into consideration the comfort of performing the gesture, the perceived fatigue, the discoverability of gestures and their learnability (ease of learning the gesture vocabulary), the simplicity of the gesture, the body part suitability, the concurrent gestures in case of intense gameplay or even the fact that gesture is performed with one hand (unimanual) or two (bimanual).

2.1.2 Interface Functionalities and Performance

According to Wachs et al. (2011, p. 62), the fact that gesture interfaces are appealing comes mainly from their flexibility and customizability. There are indeed a bunch of functionalities that can be installed when developing a new interface and that are considered as required in most of the interface types.

2.1.2.1 General Basic Requirements

Although they will probably differ depending on the application field, Wachs et al. (2011, pp. 62-64) list the basic requirements when developing a new technology as followed:

- First of all, there is a requirement for the **Price** because having better features such as improved camera quality, better framerate, greater distortion and automatic shutter speed allows for superior performance but goes hand in hand with a higher cost. Obviously, there exist some inexpensive possibilities to replace stereo cameras for achieving 3D reconstruction (for instance, using various angles to flash IR LED illuminators). However, cumulating discrete hardware components represents a significant amount of money for the average consumer, as well as for the manufacturer.

- A second requirement is the **Responsiveness** of the system as it must be able to achieve real-time gesture recognition.

- A third one is **User adaptability and feedback**, which must be understood as the ability of the system to check gesture accuracy. Two types of systems currently exist. On the one hand, are systems that are only able to detect a fixed number of pre-selected gestures, the selection being made by the system designer. On the other hand, are systems able to "*adapt to a nuanced spectrum of user selected gestures*" (Wachs et al., 2011, p.62).

- Then, a fourth requirement is about the notion of learning, in other words, the **Learnability** of gestures. The gesture vocabulary selected to command applications must be based on the ease of performing as well as remembering, two factors that are in strong association with the rate of learning.

- Afterwards, the fifth requirement is **Accuracy** that includes the detection, the tracking and the recognition of gestures. Where detection means whether or not the camera detects a hand; tracking stands for the ability to follow the hand from one frame to the next; and recognition is based on the proximity of the hand trajectories to the learned

patterns, based on distance measurements, and signals what is the probability that a given gesture could be confused with other gestures.

- Thereafter, a sixth one is about ensuring a **Low mental load**. User's mental load is likely to be added as users have to remember gesture trajectories, finger configurations and their related actions. Moreover, if the user's hands are covering the display during the use, preventing them from seeing what is happening, this also adds mental load (and physical in this case as they will probably have to redo some actions).

- Thereupon comes again **Intuitiveness** which has a clear link with other terms such as memorability and learnability (as defined above). As a reminder, the three main characteristics described by Koutsabasis and Vogiatzidakis (2019, p. 1755). It means that the interface developers, when choosing gesture types, must be careful to select gestures that have an obvious cognitive association.

- In addition, another important requirement is **Comfort**. As a matter of fact, the "Gorilla arm" syndrome, which corresponds to a situation where intense muscle tension is required over long periods to perform the gestures, should be avoided. Awkward and recurrent postures can indeed cause pressure within the carpal tunnel by straining tissues. Thus, gestures must be short and convenient while minimising stress on the hand. This muscular stress can be static if a certain posture should be maintained for a fixed amount of time, or dynamic if the hand should be moved through a trajectory.

- Furthermore, there is a requirement regarding **Lexicon size and multi-hand systems**. Some sign languages (like the American Sign Language for instance) possess a wide vocabulary with both single and two-handed gestures. Therefore, hand-gesture-recognition systems have to be developed in a way that can acknowledge this extensive lexicon.

- The next requirement is called "**Come as you are**" and is mainly the reason why I focus on cameras as communication interfaces. "Come as you are" indicates that HCI design must be such that the user does need to wear any devices such as markers, gloves or long sleeves, nor does it need to set the background or select particular lighting. The

idea behind this is to avoid the user being encumbered by any device during interaction. Consequently, the computer-vision system must be able to recognise hand gestures without the user having to wear additional aids or be connected to a device.

- As various types of users use hand-gesture systems and there is no "one size fits all" for developing related interfaces, another requirement is **Reconfigurability.** Some of the most common characteristics that vary by user type are location, anthropometric characteristics, and type and number of gestures.

- The antepenultimate requirement is the **Interaction space**. This comes from the fact that users are not standing in a fixed location with outstretched hands restricted by an envelope of virtual interaction in which gestures are recognised as assumed by most systems. For instance, this interaction envelope can surround exclusively the mobile device in the case of a mobile ubiquitous hand-gesture-recognition system.

- The second to last requirement concerns **Gesture spotting and the immersion syndrome**. When users interact at the same time with other people and devices or simply rest their hands, involuntary gestures often occur against the user's wishes. That's what we called the immersion-syndrome phenomenon. Therefore, gesture spotting is important to differentiate useful gestures from unintentional movement.

- Finally, the last requirement relates to **Ubiquity and wearability** and only refers to mobile hand-gestures interfaces. These two needs should be integrated into all aspects of daily activity, anywhere and in any context. As an example, small cameras could be fixed on the body or distributed and networked sensors could be employed when the user is in movement.

2.1.2.2 General Individual Level Computing Domains Requirements

Yong (2014, pp. 65-69) depicts the general computing domains at an individual level in the following manner, Productivity, Media Consumption, Communication and Augmentation; and lists the input and output needs for each of these categories based on a Kano analysis.

The Kano model aims to categorise products or services' features based on Herzberg's (1959)[7] (quoted by Min et al, 2018, p. 3) motivator-hygiene theory, where motivational attributes increase user satisfaction and hygiene attributes create user dissatisfaction. As an extension of this theory, the features are sorted out into six categories in the Kano model: must-be, one-dimensional, attractive, indifferent, questionable and reversal (Min et al., 2018, p. 3). In other words, the main goal of the Kano model is to classify the user's requirements regarding the effect on their satisfaction.

Min et al. (2018, p. 3) defines these six categories directly from Herzberg et al.'s book. Firstly, **must-be** features are attributes whose high level of performance does not bring satisfaction to the user, but which, if their performance deteriorates, cause dissatisfaction. They are in contrast with **attractive** features because these attributes bring a great customer satisfaction with a high level of performance but do not disappoint them in case of weak performance. In fact, attractive attributes tend to become must-be ones because customers are excited by discovering new features or characteristics of certain products or services at the beginning, but then take these as granted. Afterwards, **one-dimensional** features increase customer satisfaction when their performance increases as well. Then, **indifferent** features neither impact customer satisfaction nor customer dissatisfaction whatever their performance are, whereas **questionable** features are unclassifiable because both satisfaction and dissatisfaction can arise from their level of performance. Ultimately, **reversal** features are in opposition to one-dimensional ones because the more they perform well the more they generate customer dissatisfaction, and the more they perform badly the more they increase customer satisfaction. However, Yong (2014, pp. 65-69) only uses must-be, one-dimensional, attractive and indifferent attributes in his thesis.

➢ **Productivity**

In this computing domain mostly driven by economic purposes, users generate a significant amount of information. This large amount of information can be created by several actions: write emails, edit and publish, make spreadsheets, gather information, design and draw, create computer models and code. Generally, the usage pattern is defined by these four characteristics:

[7] Herzberg, F., Mausner, B., & Snyderman, B. B. (1959). *The Motivation to Work* (2nd éd.). John Wiley & Sons.

- The session's use duration is considered as medium to long (from thirty minutes to two hours).
- Low frequency of use per day (for instance, working eight hour a day may result in only four sessions).
- Complex data is used to generate significant amounts of information.
- Usually in a fixed setting.

According to Yong (2014, p.65) who derived them from usage patterns and sorted them out based on Kano analysis, the requirements for the productivity domain are the following. In terms of **input**, the **must-be** attributes are non-fatigue, accuracy, and responsiveness; the **one-dimensional** attribute is speed (flow); the **attractive** attribute is intuitiveness (for example, an acceptable learning cost for an efficient tool) and there is no **indifferent** attribute. In terms of **output**, the **must-be** characteristic is at least a medium size (more than 10") for the screen; the **one-dimensional** characteristic is the fact the device is portable (form factor); the **attractive** characteristics are the sound and the tactile aspect (which enhance input); and the **indifferent** characteristic is the fact that the screen size exceeds the field of vision.

➢ **Media Consumption**

In contrast with Productivity, here the users mainly consume information instead of creating it. Thus, the interaction is quite basic and is principally for navigation (read news, browse photos, watch videos, etc.). In this field, the four characteristics that defined the usage pattern are:

- The session's use duration is considered as short to medium (from five to forty-five minutes) and depends on the consumed media.
- High frequency of daily use.
- A simple input is used to navigate and large amounts of information are consumed.
- Usually in a fixed setting.

Thus, Yong (2014, p.66) explains that the HCI requirements are the following. In terms of **input**, the **must-be** attribute is the intuitiveness (it is unacceptable to have high learning cost); the **one-dimensional** attribute is the responsiveness; the **attractive** one is the accuracy; and the **indifferent** features are the speed and the non-fatigue. In terms of output, the **must-be** characteristic is at least a small size (more than 2") for the screen; the **one-dimensional**

characteristics are the fact the device is portable (form factor), the sound and the screen size in general; the **attractive** characteristic is the tactile aspect (which enhance interactivity); and the **indifferent** characteristic is also the fact that the screen size exceeds the field of vision.

➢ **Communication**

This computing domain represents the situation in which users mainly exchange information with external parties (instant messages, calls, and videoconferences), "*resulting in rapid bursts of small pieces of information being consumed and produced at the same time*" (Yong, 2014, p.66), exchange of information having increased significantly during the recent COVID-19 pandemic with the increase in teleworking and video conferencing. For the Communication domain, the four characteristics that defined the usage pattern are:

- The session's use duration is considered as very short to short (few seconds to few minutes), each message or exchange of information being considered as one session.
- High frequency of daily use.
- Information is produced and consumed in bite-sized pieces.
- Can be in a fixed as well as in a mobile setting.

Therefore, the requirements for the Communication domain are as follows according to Yong (2014, p. 66). In terms of **input**, the **must-be** attribute is the responsiveness; the **one-dimensional** attributes are the accuracy and the speed (flow); the **attractive** ones are the non-fatigue and the intuitiveness; and there is no **indifferent** feature. In terms of **output**, the **must-be** characteristics are at least a small size (more than 2") for the screen and the fact the device is portable (form factor); there is no **one-dimensional** characteristic; the **attractive** characteristics are the sound and the tactile aspect (which enhance input); and the **indifferent** characteristic is the fact that the screen size exceeds the focus area.

➢ **Augmentation**

Finally, the last computing domain is augmentation which corresponds to the situation where users consume and produce some pieces of information which results in an augmentation of themselves in the immediate context. In other words, it is the specific use case and situation in which users expand their capability. For instance, the mere fact that a user wears a watch reinforces it by giving him or her the ability to know the exact local time. A more complex

example would be a GPS guidance system, as it augments users by giving them the ability to know their location and how to get somewhere. The augmentation usage pattern is defined by these four characteristics:

- The session's use duration is considered as very short to short (few seconds to one minute). Nevertheless, some augmentations are activated for long periods of time (as in the case of the GPS on a long trip), but the real time that attracts users' attention is only a few seconds at a time.
- Very high frequency of daily use.
- Sensor's data and limited user input for the most part produce or consume information.

Yong (2014, p.67) describes that the HCI requirements of the augmentation domain are the following. For the **input**, the **must-be** attributes are the intuitiveness and the responsiveness; the **one-dimensional** feature is the accuracy; the **attractive** one is the non-fatigue; and the **indifferent** attribute is the speed (flow). For the **output**, the **must-be** characteristics are the appropriate contextuality (screen/led/sound/tactile), and the fact the device is portable (form factor); there is no **one-dimensional** one; the **attractive** characteristic is the fact the device is wearable (form factor); and the **indifferent** attribute is anything that is not in the context used.

Listing all these requirements allowed Yong (2014, pp 67-68) to extrapolate them into the four quadrants of individual computing which is based on the complexity of the interaction as well as the duration of use. Moreover, he also created a more detailed map of the requirements (2014, p. 69).

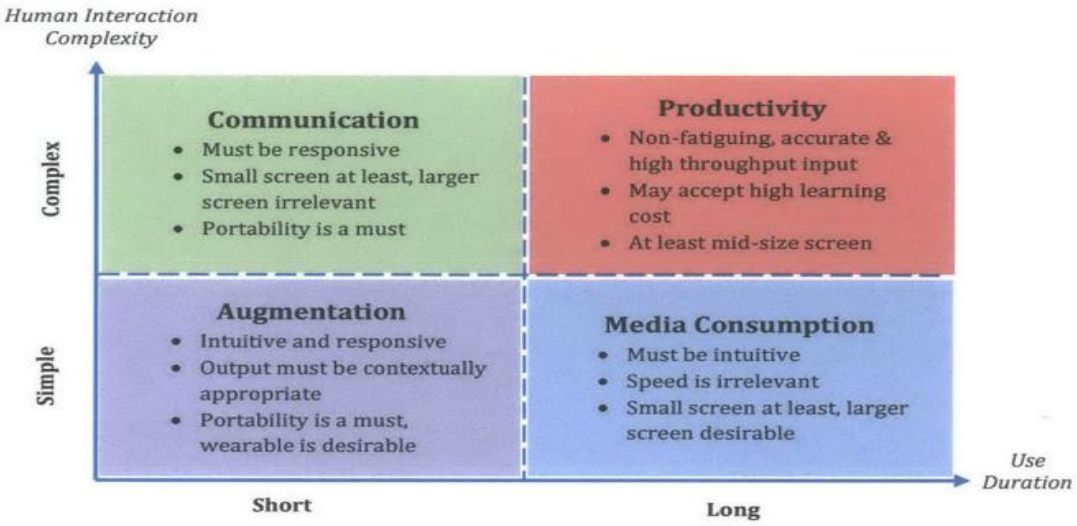


Figure 2- "Four quadrants of computing domains and corresponding HC1 characteristics" (Yong, 2014, p.68)

	Desktop/ Workstation	Laptop	Tablet	Smart phone	Wearable
Main Activity	Productivity	Productivity	Media consumption	Communication	Augmentation
Ergonomics	Completely stationary	Portable but stationary usage	Portable and mobile usage	Portable and mobile usage	Complete mobile
Screen Size	Large (27" – 15")	Large (15" – 10")	Medium (10" – 5")	Small (5" – 2")	Very small – None (<2")
Interaction Complexity	Very complex, fully customizable	Complex, customizable	Limited	Limited	Very limited
Input style	Keyboard, mouse	Keyboard, mouse	Fingers, voice	Fingers, voice	Fingers, voice, sensor data
Input Performance	Precise, fast	Precise, fast	Less precise, slower	Less precise, slower	Precise, fast
Input feedback	Visual + Tactile feedback	Visual + Tactile feedback	Visual + Pseudo-tactile feedback	Visual + Pseudo-tactile feedback	Visual + tactile feedback
Computing power	> Quad cores 2 – 3GHz speed 4 – 16GB RAM	Dual – Quad cores 1.5 – 2.5GHz speed 2 – 8GB RAM	One – Quad cores 0.9 – 2Ghz speed 0.25 – 2GB RAM	One – Dual cores 0.7 – 2 GHz speed 0.25 – 2GB RAM	One core / ASIC < 0.5GHz speed < 0.5GB RAM
Use Frequency	Infrequent (3 –4 times a day)	Infrequent (3 – 4 times a day)	Frequent (5 – 10 times a day)	Often (10 – 20 times a day)	Very often (>15 times a day)
Use Duration	Long (0.5 – 2 hours per session)	Long (0.5 – 2 hours per session)	Medium (10 – 30 mins per session)	Short – medium (1 – 30 mins per session)	Very short (within 1 min per session)

Figure 3- "Mapping of form factors to individual computing activities" (Yong, 2014, p.69)

2.1.2.3 General Collaborative Level Computing Domains Requirements

Concerning the collaborative level, Yong (2014, pp. 70-71) only explains the requirements for the productivity and the media consumption domains. This seems quite logical given that we rarely communicate in a collaborative way (for example, it certainly happens that sometimes a message is written by several people but in the end it is only one user who sends it). Moreover, when a technology increases our capacities, it is also mostly done individually (in the case of GPS in cars, the increase concerns the driver rather than the passengers).

- **Collaborative productivity**

According to Yong (2014, p. 70), Collaborative productivity happens when multiple parties are working in collaboration simultaneously to produce, consume or exchange information. Typically, there are three forms of interactions between the participants: one to many, many to one, and many to many. One to many interactions happen for instance when someone presents something to a group or trains a group. Many to one happen for example when an employee provides a meeting report to its manager. Finally, Many to many happen when a group of people discuss, brainstorm or even take part in a decision making process. Two types of collaboration exist depending on the locations of the participant, the local and the online one. At the time of Yong's thesis, local collaboration was already being rapidly replaced by online

collaboration, and it is clear that this situation has not changed, quite the contrary. In order to combat the Covid-19 pandemic, teleworking has been made compulsory in most countries in occupations that allow it. This has further increased online collaboration. However, Yong (2014, p. 70) explains that the main problems were a lack of bandwidth, which includes creative tools, emotional signals and contextual information, as well as a lack of frequency of communication compared to local collaboration, "*as suggested by the Allen Curve created by MIT researcher, Thomas Allen*". This is confirmed because from my point of view as a student, the fact that I had to go into a completely online mode to do group work for example was not at all optimal from a collaborative point of view. Indeed, a lot of problems arose when we wanted to work separately on the same online shared document for example. Often, the shared document tended to crash completely and we had to try to arrange not to work on the document at the same time, which is not optimal from a collaborative point of view. In any case, whether it be for online or local collaboration, here are the requirements related to IT environments for effective collaboration according to Yong (2014, p.70):

- First of all, in order to support the interactions between participants, the communication bandwidth should be fairly high.
- Secondly, there is a need to disseminate information by using mixed media support (text, graphics, video).
- Thirdly, a large synchronised public display should be installed as a visual focus, and optionally multiple screens could be used to present peripheral information.
- Fourthly, each participant's information should be easily and transparently changed in public view, and optional support for a large number of device classes should be provided.
- Finally, as creative tools, free-form media support (e.g. whiteboard, sketch, etc.) must be provided optionally, in analogue or digital form.

To sum up, collaborating to produce information requires at a minimum the same level of interaction complexity as the individual level, and a large screen or multiple screens as an output footprint.

➢ **Media consumption in groups**

The difference between group media consumption and individual media consumption is only that multiple parties are consuming simultaneously. Apart from that, they are relatively equivalent. The observations are exactly the same as for collaborative productivity, but a larger screen size is required than at the individual level.

2.1.2.4 Man-Machine Dialogue Requirements

In his book, Landragin (2013, p. 17) describes the different functionalities for recording devices using Man-Machine Dialogue, although this list is non-exhaustive, it includes the main characteristics that a device should possess. A recording device can includes : speech; lip-reading (to help user and particularly useful in noisy environment, with disabled user, or when whispering as it could replace speech recognition); user recognition; face location and face and mouth or eye tracking, and in case of eye tracking also eye direction (on the one hand to oversee attention in relation to the dialogue, and on the other to help understand the reference to an object present in the surroundings); facial emotion; gesture pointing (in particular those of the hand but also general gestures performed with any part of the body); and finally force-feedback gesture for haptic interaction (device managing the fact the user could generate a potential resistance, while recording hand's position). Nevertheless, Landragin (2013, p. 17) explains that there is no perfect system that includes all these recording requirements at the same time and in real time.

2.2 Existing gesture classifications

Currently, a large number of gesture classifications have been created and used to sort the gestures acquired in gesture elicitation studies. The aim of this section is therefore to make a non-exhaustive chronological summary (in view of the number of existing classifications) by grouping together as far as possible the main classifications and measuring the pros and cons of each of them. In general, people tend to base their classifications on these main classifications while adapting them to the context in which they find themselves (type of device, type of user, field of application, etc.). Therefore, first and foremost the origins of the first classifications will be explained shortly and then the most recent ones will be described in detail.

2.2.1 Efron

The first gesture classification was proposed by Efron in 1941[8] (Vafaie, 2013, p. 4) and adapted by himself in 1972 [9](Aigner et al., 2012). This first classification aimed to analyse the conversational gesture behaviour of southern Italians and Jewish people during their immigration in New York City. According to Efron (1941) quoted in Vafaei (2013, p. 4), there was two different classes of gestures: the gestures that had a meaning independent of speech were considered as "Objective"; and the ones linked to speech were considered as "Logical-discursive". The Objective gesture category was divided in three sub-categories: Deictic, Physiographic, and Symbolic (or Emblematic); while the "Logical-discursive" gesture class was divided in two sub-categories: Batons and Ideographic.

2.2.2 Kendon

Then, even if several classifications were adapted from Efron's work, the second important early classification was developed by Kendon in 1981[10] (Vafaei, 2013, p. 5). Kendon classified the gestures as a continuum that depends on formality and speech (Vafaei, 2013, p.5). Aigner et al. (2012) describes this continuum as the reflection of the accompaniment of speech by gesture. The important evolution was that Kendon replaced the typical gesture by gesticulation, and therefore proposed the following continuum with the need to support speech in order to ease communication declining in this order: Gesticulation → Language-like gestures → Pantomimes → Emblems → Sign Language (Vafaei, 2013, p. 5). Moreover, according to Aigner et al. (2012) each of these categories (excluding "pantomimes") had one or several sub-categories. As a matter of fact, Gesticulation was divided into Beat and Cohesive sub-classes, Language-like into Iconic, Emblems into Deictic, and Sign Language into Symbolic.

2.2.3 McNeill

Afterwards, based on Kendon's work, McNeill (1992)[11], quoted by Aigner et al. (2012) and Hiyadi (2016, p. 25), and reviewed by Feyereisen (1994, pp. 150-151), proposed a new

[8] Efron, D. (1941) *Gesture and Environment*. Morningside Heights, NY: King's Crown Press.
[9] Efron, D. (1972) *Gesture and environment*. The Hague: Mouton.
[10] Kendon, A., Sebeok, T. A., and Umiker-Sebeok, J. (1981) *Nonverbal Communication, Interaction, and Gesture: Selections from SEMIOTICA*. Mouton Publishers.
[11] McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought* (New edition). University of Chicago Press.

classification based on five categories of gestures, namely beat, cohesive, deictic, metaphoric, and iconic. And as some of those classes are style used in today's classification, it seems important to explain what they mean.

Beat gestures are rhythmic gestures that are in synchronisation with the speech and are in fact the gestures called "Batons" by Efron (Vafaei, 2013, p. 7). When people make these kinds of gestures, it is like they are beating the tempo of their speech with their hands or fingers.

Cohesive gestures are, according to Hiyadi (2016, p. 25), separate in time although they are thematically related because someone external has interrupted the communication during the process.

Deictic gestures are pointing actions that are used to identify a person, an object or a real location, which in this case are considered as concrete. These kinds of gestures can also be considered abstract if they are used to point at an abstract location (such as a specific location in a 3D space) or a time period (Vafaei, 2013, p. 7; and Hiyadi, 2016, p. 25).

Metaphoric gestures always reflect abstract concepts (narrative genre, knowledge, language itself, …) under the form of an image (Hiyadi, 2016, p. 25; and McNeill, 1992, quoted in Vafaei, 2013, p. 7).

Iconic gestures are similar to metaphoric ones in the sense that they also reflect concepts under the form of an image, but the main difference is that they are used to represent concrete ideas (Vafaei, 2013, p. 7). Moreover, McNeill (1992), quoted in Vafaei (2013, p. 7) explains that gesture is considered as iconic if it is closely related to the semantic content of the speech.

2.2.4 Quek et al.

Quek et al. (2002, p. 172), which includes McNeill, have brought two important terms that are **Manipulative** and **Semaphoric** to categorise gestures. They started from the continuum developed by Kendon, took into account the evolutions developed by McNeill and incorporated these two new terms.

Manipulative gestures are those whose objective is to control an entity by applying a close relationship between the actual movements of the hand or arm and the entity to be manipulated in order to take control of it.

Semaphoric gestures are "*systems of signaling using flags, lights, or arms [Brittanica.com]. By extension, we define semaphoric gestures to be any gesturing system that employs a stylized dictionary of static or dynamic hand or arm gestures*" (Quek et al., 2002, p. 172).

However, these two gesture categories suffer from major deficiencies according to Quek et al. (2002, p. 173). From one side, the main issue with manipulation is that in most of the cases, even if it is done significantly by using human natural hand gestures, it implies to handle the object being manipulated. In addition, the problem with free-hand manipulation interfaces is that they use visual feedback almost exclusively, and thus lack the feedback brought by the manipulation of the artefact. On the other side, the problem with semaphores is that they are only a small part of the use of hands in natural human communication. This lack of naturalness was already mentioned by Wexelbat (1997)[12], quoted by Quek et al. (2002, p. 173), when he reviewed the challenges to automatic gesture recognition, as he highlighted the necessity to develop systems that can recognise natural, non-posed, and non-discrete gestures.

Even if Efron, Kendon, McNeill and Quek et al., really gave birth to the first classifications and made them evolve, their classifications were only used in the framework of performed gestures in relation to the discourse. Therefore, it was important to start the extension of the gestures acquired to the whole range of the body, from head to toe in order to represent all possible types of interaction through human-machine interfaces. This part will therefore depict the various important classification that can be used in HCI.

2.2.5 Karam & Schraefel

Karam and Schraefel (2005, pp. 2-22) proposed a whole taxonomy of gestures in HCI, by categorising gestures into four different categories: application domains, enabling technologies, system responses and gesture styles.

Karam and Schraefel (2005, pp. 13-20) listed ten application domains to categorise gestures, in which they added some sub-categories. The first domain proposed is **Augmented**

[12] WEXELBLAT, A. 1997. Research challenges in gesture: Open issues and unsolved problems. In Proceedings of the International Gesture Workshop: Gesture & Sign Language in HCI, I. Wachsmuth and M. Frohlich, Eds., Springer, Bielefeld, Germany, 1–11

and Virtual Reality, with the following sub-categories: Immersive interactions and avatars; Manipulation and navigation in virtual words; Augmented reality; and Robotics and telepresence. Then, the second domain is the one of **Desktop and Tablet PC Applications**, which is divided into: Graphics and drawing applications; Computer-supported cooperative work and Multimodal interfaces. Afterwards, they described the **Pervasive and Mobile Interfaces** domain with the sub-category of Wearable computing, as well as the one of **Ubiquitous Computing and Smart Environments** which includes the sub-category of Tangible computing. Finally, they refer to some domains without considering any sub-category, such as **Three-Dimensional Displays**; **Games**; **Telematics**; **Adaptive Technology**; **Communication Interfaces**; and **Gesture Toolkits**.

Then, in their work, Karam and Schraefel (2005, pp. 9-13) proposed to classify gestures regarding the input technology that helps interaction, and use two classes of technologies, namely **Non-perceptual** and **Perceptual** input.

Perceptual inputs are considered as technologies allowing the user to use gestures to communicate with the system without having to wear, to carry or to touch any device. Thus, thanks to visual, audio or motion sensors, the perceptual input technology has the capacity to receive "*sensory input data from the user through their actions, speech or physical location within their environment"* (Karam & Schraefel, 2005, p. 12). The two major perceptual input technologies being technologies using computer vision and technologies using remote sensors.

Non-perceptual inputs, in contrast**,** are technologies necessitating the user to wear, carry or touch a device to communicate with the computer. The input gesture is transmitted through an intermediary device. Karam and Schraefel (2005, pp. 10-12) proposed seven non-perceptual input categories: Mouse and pen input; Touch and pressure input; Electronic sensing: wearable or body mounted; Electronic sensing: gloves; Electronic sensing: sensor-embedded objects and tangible interfaces; Electronic sensing: tracking devices; and audio input.

According to Karam and Schraefel (2005; pp. 20-22), the system response or output, which has a major role in the user experience and is considered as the end result of a gestural interaction, is one of the main characteristics to differentiate the various possible gesture interactions. In their work, they identify three main categories: **Visual output**, with two sub-categories depending if the output is provided in 2D or 3D; **Audio output**; and **Central Processing Unit (CPU) or Command Directed Output**.

After reviewing Quek's research as well the existing literature of the time, Karam and Schraefel (2005, pp. 4-9), considered five main categories of gestures, namely gesticulation, deictic, manipulations, semaphores, and language gestures.

The first category of gestures in Karam and Schraefel's taxonomy is **Gesticulation**. They define gesticulation as the category which includes gestures that naturally accompany speech and thus are not pre-recorded in the interface vocabulary. In other words, gesticulation does not impose the user to learn and perform a specific set of gestures different from natural speech related gestures.

The second gesture style is **Deictic** gesture. As explained before, this type of gestures correspond to pointing gestures used to identify a person, a person, or a real spatial location within the specific context of the application domain (desktop computer, virtual reality, mobile device, …). In addition, this type of gestures can be considered as concrete as well as abstract depending on what they aim to point out (Vafaei, 2013, p. 7; and Hiyadi, 2016, p. 25; and Karam & Schraefel, 2005, p. 4).

The third type of gesture used by Karam and Schraefel is **Manipulative** (or manipulations) gestures. These gestures are defined the same way as they were Quek et al. (2002), but Karam and Schraefel (2005, p. 5) add the dimensionality in which they are performed and create four kinds of subcategory. Their four sub-categories include gestures: with "*two degrees of freedom for two-Dimensional interactions*"; with "*multiple degrees of freedom for two-dimensional interactions*"; "*with tangible objects for three-dimensional interactions";* and finally "*for real-world physical object interactions*" (Karam & Schraefel, 2005, pp. 5-6). Therefore, manipulative gestures are classified regarding their degree of freedom as well as their dimensionality (2D or 3D).

The fourth class of gesture in Karam and Schraefel's categorisation is **Semaphoric** (or semaphores) gestures, that are also defined as they were by Quek et al. (2002). However, three sub-categories are considered by Karam and Schraefel (2005, p. 7): **Static, Dynamic** and **Stroke** gestures. As already explained by Quek et al. (2002, p. 172), semaphoric gestures occur in a system using "*a stylized dictionary of static or dynamic hand or arm gestures"*, and Karam and Schraefel (2005, p. 7) confirm this by explaining that semaphoric gestures can "*involve*

static poses or dynamic movements". Moreover, they bring the notion of stroke gestures, a type of gestures made with a physical device (a mouse or a stylus) to mimic strokes or marks corresponding to various interface commands.

Last but not least, **Language** gestures correspond to gestures used to represent sign languages and which are most of the time independently considered because they are based on linguistic and "*are performed using a series of individual signs or gestures that combine to form grammatical structures for conversational style interfaces*" (Karam & Schraefel, 2005, p. 8). We could think they are considered as gesticulation but what makes them different from gesticulation is the fact they are pre-recorded in the interface recognition system.

2.2.6 Aigner et al.

Aigner et al. (2012), after having based a pilot study on Karam and Schraefel's taxonomy (2005), needed to modify it slightly as the gesticulation class did not allow them to distinguish between gestures they considered iconic and those they considered pantomimic. Moreover, they found that the subcategories of static and dynamic gestures could also be applied to iconic gestures, and not only to semaphoric ones. Nevertheless, there is no possibility to perform strokes with iconic gestures. In addition, they also chose to modify deictic by pointing because they did not use communication speech in their application. Therefore, as some modifications have been made from Karam and Schraefel (2005) taxonomy, below is how Aigner et al. define each of the terms in their taxonomy, as well as a visual representation of their classification with an example for each gesture type.

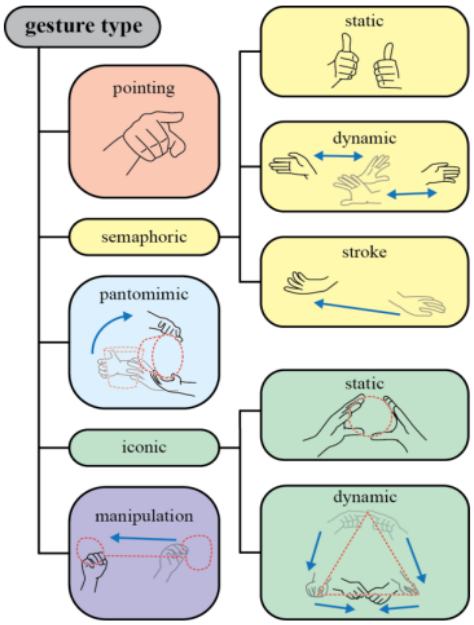


Figure 4- Gesture mapping (Aigner et al., 2012)

Pointing gestures are used as the indication of objects and directions. However, this type of gesture is not necessarily performed with a stretched index finger (as in the representation above), but can also be done by another finger (generally thumb), a flat palm, multiple fingers, …

Semaphoric gestures "*are used to convey specific meanings*" (Aigner et al., 2012) and correspond to hand posture or movements. Generally, the gesture performed is not related to the meaning and is thus strictly learned. Therefore, semaphoric gestures depend, from Aigner et al.'s point of view, on the experience and background of the user.

Static semaphoric refers to a specific hand posture. For instance, "okay" being represented by a thumbs-up.

Dynamic semaphoric transfer information through their temporal aspects. For example, "rotation" represented by a circular hand motion.

Semaphoric stroke corresponds to hand flicks and are equivalent to dynamics as they are identified by hand movement only, "*but which are single, stroke-like movements"* (Aigner et al., 2012). For instance, a simple sideways stroke of the hand, meaning "move that object away".

Pantomimic gestures represent the situation where the user performs or imitates a particular task, which often requires specific hand postures and movement. The user doing pantomimic gestures does not use any physical objects in most of the cases but uses imaginary

items. An example would be the one of an actor grabbing an imaginary glass, filling it with water and then releasing it. They usually consist of a sequence of low-level gestures.

Iconic gestures are performed to provide information like particular sizes, shapes, movement paths about items or entities.

Static iconic gestures refer to a user performing static hand postures, but which do not have a particular meaning in the pre-defined vocabulary of the interface as in the case of a static semaphoric. On the diagram above, the user is making an "O" with both index and thumb to mean a "circle".

Dynamic iconic gestures are mostly performed in order to represent paths or shapes, like making circles with hand to mean "circle". The main difference with semaphoric strokes is, when performing concatenated flicks, movements are generally performed more slowly. Moreover, with dynamic iconic gestures, the range of motion has meaning, unlike strokes.

Manipulation gestures are performed in short feedback loops in order to lead the motion. A close relationship between the movements of the user and the movements of the object being manipulated is therefore present. The presence of this feedback loop allows to distinguish manipulation gestures from pantomimic and dynamic iconic ones. As a matter of fact, when performing manipulation gestures, users tend to make sure the object "follows" before going on, instead of executing it before, causing a reaction only thereafter.

2.2.7 Vafaei

Vafaei (2013) proposed in his thesis a new taxonomy that is mainly based on two taxonomies: the one of Wobbrock et al. (2009)[13] for surface gestures; and the one of Ruiz et al. (2011)[14] for 3D motion gestures. Vafaei's (2013, p. 13-32) taxonomy is composed of two main groups: Gesture mapping which corresponds to how users tend to map gestures with tasks; and physical characteristics which help to define gesture's physical attributes.

[13] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. (2009). "User-defined gestures for surface computing," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA pp. 1083–1092.

[14] J. Ruiz, Y. Li, and E. Lank. (2011). "User-defined motion gestures for mobile interaction," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA. pp. 197–206.

2.2.7.1 Gesture mapping

Gesture mapping is composed of five main dimensions that themselves possess several sub-sections: *Nature* (Manipulative, Pantomimic, Symbolic, Pointing and Abstract); *Form* (Static, Dynamic, Stroke); *Binding* (Object-centric, World-dependent, and World-independent); *Temporal* (Continuous and Discrete); and finally, *Context* (In-context and No-context).

Nature dimension corresponds to Vafaei's classification highest level of abstraction and aims to map gestures to the intended task and/or the entities and describe how they relate to one another. As **Manipulative**, **Pantomimic** and **Pointing** gestures have already been defined in Aigner et al.'s (2012) classification, it is not necessary to define them again. Moreover, **Symbolic** gestures correspond to the Semaphoric gestures explained above also in Aigner et al.'s (2012) study, and only apply to hand and arm gestures in this taxonomy. As a matter of fact, Vafaei (2013, p. 18) considered that even if foot or head can be used to represent a symbol, few users are going to do it. Therefore, the only subsection that must be defined is the **Abstract** nature of gestures. According to Vafaei (2013, p. 18), the nature of a gesture is considered as Abstract if the gesture mapping is arbitrary, and this type of gestures is usually used in case of "*limited space of available gesture*".

Then, the **Form** dimension contains the same subsections as the one of Semaphoric gestures in Aigner et al. (2012) classification. In other words, this dimension aims to sort gestures regarding if they include movements (**dynamic**), flicks (**stroke**) or no motion (**static**).

Afterwards, the **Binding** dimension helps to know if specific information is required to perform the gesture, whether about the affected or produced object, or about the application's environment. Furthermore, Vafaei (2013, p. 21) points out that this dimension only represents the attribute of the gesture and not the task.

Object-Centric gestures have only one requirement which is to know the object location (that is itself defined in relation to the object). For instance, an object-centric gesture would be touching or pointing to select a specific location on a computer screen.

World-Dependent gestures correspond to the ones needing to acquire information about the application environment (world) like "*tapping in the top-right corner of the display or dragging an object off-screen*" (2013, p. 20).

World-Independent gestures are those that do no require any information about the object affected or produced, or about the world, and which can be performed usually on any space of the surface (in case of surface device) or anywhere in the air (in case of mid-air gestures).

As for the **Temporal** dimension, it addresses the relationship between gesture recognition and its impact. On the one hand, if the gesture impact happens after the gesture is performed, therefore the gesture is categorised as **Discrete**. On the other hand, if the gesture recognition is carried out by the computer as a continuous process, simultaneously with the gesture being made, and the impact can be seen as "in progress", the temporal dimension is considered **Continuous**. According to Vafaei (2013, p. 22), because of the feedback loop between human and computer, manipulative gestures are usually continuous.

Finally, the **Context** dimension specifies whether the context is necessary to figure out the meaning of the gesture. When the user employs the same gesture for two or more various tasks/goals, the gesture is considered "**In-Context**" as the meaning clearly depends on the context. If not, the gesture is considered as "**No-context**".

2.2.7.2 Physical characteristics

In Vafaei's taxonomy (2013, pp. 23-32), physical characteristics refer to physical attributes that can be given to the gestures themselves. There are six characteristics: Dimensionality, Complexity, Body Part, Handedness, Hand Shape, and Range of Motion, which each contains one or several subcategories.

Dimensionality must be understood as the number of axes engaged in the motion, and therefore contains four subcategories: **Single-Axis** correspond to a movement performed in only of the X,Y or Z coordinates (either on the surface or in the air); **Double-Axis** gestures always happen on a 2D surface (for instance in tabletop application); **Tri-Axis** gestures occur when a translation or a rotation is performed as gesture in 3D space; and **Six-Axis** gestures occur when the gesture performed includes both rotation and translation in 3D space.

The **Complexity** characteristic defines whether "*the gesture consists of a single Simple gesture, or includes multiple Simple gestures*" (Vafaei, 2013, p. 25). A single movement or

posture performed without repeating, spatial discontinuities such as inflection points, and breaks in the middle, is considered as a **Simple** (atomic) gesture. Therefore, if we can decompose a gesture into two or several Simple (atomic) gestures, this gesture is considered as a **Compound** gesture.

The **Body Part** characteristic speaks for itself and describes the body part(s) that is (are) used to perform the gesture. In his taxonomy, Vafaei (2013, p. 25) considered mainly **arms** and **hands**, but also took into consideration **head**, **shoulders**, and **foot**.

The **Handedness** dimension only stands for hand and arm gestures and contains three subsections. This dimension is considered as: **Dominant** if the user used his/her dominant hand or arm to perform the gesture; **Non-Dominant** if the user used his/her non-dominant hand or arm to perform the gesture; and **Bi-manual** if both hands/arms are used by the user to perform the gesture.

Concerning the **Hand Shape** dimension, which also only stands for arm and hand gestures, is one of the key characteristic of Vafaei's (2013) taxonomy, and correspond to "*the configuration, form and posture of fingers while the gesture is performed*" (Vafaei, 2013, pp. 26-27). To make this dimension more visual; Vafaei (2013, pp. 28-29) has included in his work a "*set of common hand shapes among users for gestural interaction with computers*" and named them, using American Sign Language to name some of them.

TIM Open	Bent Hand	Free Hand	Curved Hand	Claw
Flat Hand	Ball Holding	Joined Fingers	Index Finger	Snapping Fingers
Thumb Up	Thumb Down	Fist	Open Hand	Pinch
ASL-L	Free ASL-L	ASL-U	Free ASL-U	ASL-Q
ASL-F	ASL-X	ASL-V	ASL-C	

Table 1- Most common hand shapes (Vafaei, pp. 28-29)

Last but not least, the **Range of Motion** (ROM) is the other key characteristic of Vafaei's (2013) taxonomy. However, to understand what it describes, some context must be provided in terms of human anatomy. In this context, ROM is defined as: "*the capability of a joint to go through its complete spectrum of movements*" (Physiopedia, 2022); and a joint is defined as: "*a form of connection between bones*" (Physiopedia, 2022). In addition, Vafaei (2013, p. 30) explains that the Expected ROM corresponds to the "normal ROM" of each joint in the human skeleton.

Now let's see how this concept is applied in this taxonomy. Range of Motion is defined as:

the ratio of the movement of the corresponding joint, which is attached to and is in charge of moving the involved body part of the gesture, over the Expected ROM of that particular joint and has two values: Small (less than 50% of the Expected ROM of the joint) and Large (50% or more of the Expected ROM of the joint). (Vafaei, 2013, p. 31)

Therefore, for any gesture, it is necessary to first determine the body part involved and the corresponding joint, in order to assess, as **Small** or **Large**, the ROM in the gesture classification.

2.2.8 Hiyadi

In her PhD thesis, Hiyadi (2016, pp. 24-26) proposes a new taxonomy to categorise gestures, that nevertheless focuses only on dynamic upper body gestures, whether simple or compound as defined by Vafaei (2013). This taxonomy was mainly based on Ottenheimer (2012)[15] (from emblems to adaptors) and McNeill (1992) (subcategories of illustrators).

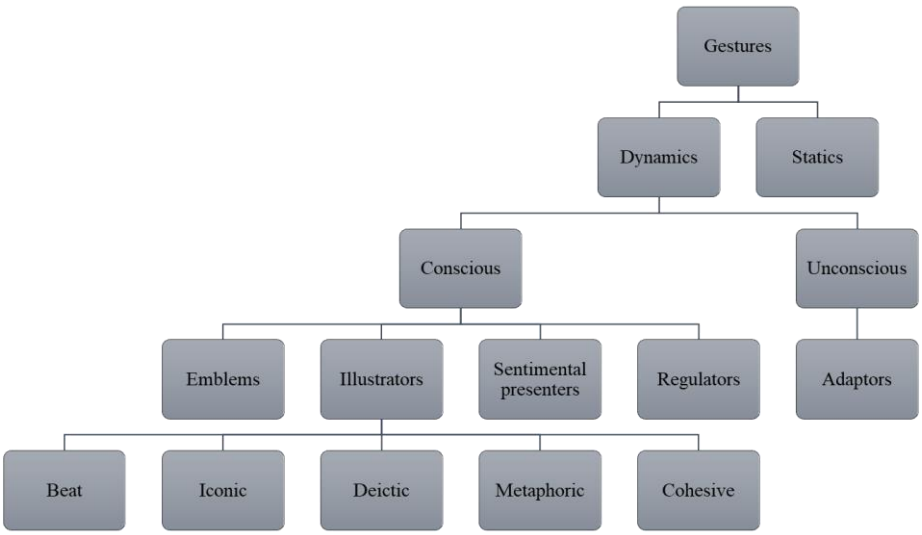


Figure 5- A taxonomy for gestures (Hiyadi, 2016, p.26) [My translation]

Therefore, as already explained several times, gestures can either be **dynamic** or **static**, but Hiyadi (2016) only focuses on dynamic gestures. Then, dynamic gestures can be conscious

[15] Ottenheimer, H. J. (2012). The anthropology of language: An introduction to linguistic anthropology. Cengage Learning

(intended by the user) or **unconscious** (as explained in the case of immersion syndrome in section 2.1.2.1).

If the gesture is unconscious, thus it is considered as an **Adaptor**, which is a gesture used to release tension in the body, such as shaking the head (Hiyadi, 2016, p. 25). However, if the gesture is conscious, four categories of gestures can be used: emblems, illustrators, sentimental presenters and regulators.

Emblems correspond to gestures (sometimes also called emblematic) replacing the words of a short verbal communication, like saying goodbye with a hand gesture for instance. Thus, they are mainly influenced by culture.

Illustrators define gestures that are thought and speech related. That is why, Hiyadi (2016, p.25) used as subcategories the ones developed by McNeill (1992), which are as reminder: Beat, Iconic, Deictic, Metaphoric, and Cohesive gestures.

Sentimental presenters are gestures that give away the emotion or the intentions of the communicator when they are executed. Therefore, they are not really influenced by the culture.

By the end, **Regulators** aim to regulate the interaction itself. An example would be the one of gestures that helps to regulate speaking in a conversation to avoid people cutting each other off.

2.3 Gesture information representation

After presenting how to collect gestures thanks to GES, and how to classify them, let's see how we can represent the gesture information that occurs between humans and systems in HCI. This section explains two ways of representing gesture information according to McAweeney et al. (2018) and Bernsen and Dybkjær (2009). Using graphics and illustrations to represent gestures provides a rich vocabulary of design elements for expressing the dimensions of interaction

2.3.1 McAweeney et al.

The aim of the McAweeney et al. (2018) study was to analyse the interpretation, implementation, and refinement of gesture representation by users through adaptation using techniques from recent GES. Gestures are mostly represented by graphics and illustrations, which is useful as a rich vocabulary of design elements is provided thanks to that in order to express the dimensions of interaction (McAweeney et al., 2018, p. 1). However, according to

McAweeney et al. (2018, pp. 1-2) although more and more gesture-based interfaces are developed, and despite that the most popular gesture-based technologies possess interface design guidelines, there is currently few single graphical vocabulary accepted conventionally to articulate gesture's dimensions. Therefore, the existing guidelines mentioned, as well as academic papers, are using disparate design to represent gestures dimensions which result in a low guarantee for the correct gesture's interpretation either by potential end-users, other researchers, or system developers. This problem usually arises because researchers and designers, who have full knowledge of the implementation of the gesture, design representations by making assumptions about what should or should not be communicated in the representation, which leads to omissions. Moreover, McAweeney et al. (2018, p. 2) point out another problem which consists of a breach in elicitation research as existing studies focused principally on the gestural input form of the user, and not on how users understand the communication of gestures as symbolic output. Thus, in order to accelerate the learning speed of a system, recall as well as task's execution, and to reduce the error rate, this gap between computer and user should be narrowed.

McAweeney et al. (2018, p. 4) therefore analysed GES in order to propose a "*taxonomy of design elements used to communicate gestural interactions*". This taxonomy classifies gestures into six dimensions, respectively Body context, Environmental context, Perspective, Frame, Colour, and Gesture Elements, which all together contain twenty-six (sub)categories.

Body context includes those parts of the body that are extraneous to the representations and are not vital to the performance of the gesture. For instance, representing the whole-body for a gesture requiring only a thumb-up. There are four various categories: **Full-body**, **upper-body**, **lower-body** and **other body parts**.

Environmental context, in the same way, represents objects external to the body and not intrinsically part of the gesture. An example would be the one of a TV represented in the case of a hand swipe gesture to move to the next channel. This dimension is divided into two categories: **physical objects** and **virtual objects**.

Perspective dimension contains five categories encompassing the representation's point of view. This includes on the one hand the angle from which the representation is represented (**Mirror, Bird's eye, Side angle**). On the other hand, the way in which the

representation is presented to the user, either read as if he/she were performing it himself/herself (**1st person**) or as if the gesture were shown to him/her (**3rd person**).

The **Frame** indicates the number of frames used to communicate with the machine. There are two possibilities: either the gesture was made with other gesture elements to communicate aspects such as movement and time but using only one frame, in which case it is considered "**Single**"; or a single gesture was communicated using multiple frames, in which case it is considered "**Multi**".

Colour simply "*includes the use of both hue and shade for encoding gestural aspects*" (McAweeney et al., 2018, p. 4). Only two possibilities exist: **Yes** or **No**. In addition, with Perspective and Frame, the Colour dimension represents the necessary structural elements for the design of any representation.

Gesture elements, for their part, grouped with Body and Environmental contexts, represent representation details, and are added to structural elements to better communicate gesture's aspects. This dimension regroups everything that is added to facilitate the representation, like symbols and signs. It contains eleven categories: **1-Sided arrow**; **2-Sided arrow**; **Dotted lines**; **Ghost**; **Finger trail**; **Other motion lines**; **Touchpoints**; **Numbers**; **Text**; **Bending Joints**; and **Axis.**

2.3.2 Bernsen and Dybkjær

First of all, it is important to define what is a unimodal interactive system as well as a multimodal one, but also what is a modality. According to Bernsen and Dybkjær (2009, p. 72), a unimodal interactive system is composed of an input and an output that are used with the same single modality, while a multimodal system has also an input and an output but which are used with two different modalities for at least one or both of them. Concerning modality, it is defined as "*a particular way of representing information in a particular medium*" (Bernsen & Dybkjær, 2009, p. 72). Bernsen and Dybkjær (2009, pp. 72-74) proposed a taxonomy of unimodal modalities for three types of media related to the human senses: graphics (vision), haptics (touch), and acoustics (hearing). Nevertheless, in order for the table below to be useful as a toolbox of input and output modalities when building multimodal systems, they have carefully compressed the less relevant modalities of the time. But, before going through this

table, some explanations must be given about its structure in order to understand all the abbreviations as well as the different levels.

At the *super level*, Bernsen and Dybkjær (2009, p. 75) classified modalities into four categories: Linguistic; Analogue; Arbitrary; and Explicit Structure modalities. *Linguistic* modalities are the language-based means of representing information; *Analogue* modalities are the similarity-based, between the representation and what is being represented, means of representing information. An example would be images that resemble what they are images of; *Arbitrary* modalities are used to give ad hoc meaning for a certain purpose by means of an element that has no specific meaning in advance. For example, using an element to highlight text to be discussed or deleted; finally *Explicit structure* modalities are "*something we insert to mark representational structure, like the line grid in common tables*" (Bernsen & Dybkjær (2009, p. 75). The following levels of the table were obtained by inheritance using a common mechanism. As a matter of fact, Bernsen and Dybkjær (2009, p. 75) explain that each level is based on its predecessor adding distinctions in order to subdivide the modality cited at the antecedent level into different more specific modalities. Therefore, the table is composed of four main levels: super level; generic level; atomic level; and subatomic level.

Bernsen & Dybkjær (2009, pp. 76-79) explains the different modalities we can find in the *generic level*, which is the most important in this taxonomy. It regroups twenty generic modalities that possess a standard meaning in the sense of a meaning shared by some (sub-)culture. Shared meaning is essential for communicative interaction, as it allows us to represent information in a certain modality in such a way that all members of our (sub-)culture understand it. However, some shared meaning could correspond to non-standard meaning, which is the opposite. Non-standard meaning function generally comes from standard meaning words that were used in some specific cases such as metaphors or metonymies, which were therefore understood at a certain period but which lose their meaning over time, without us even realising it.

Then, the term static and dynamic are not used in their taxonomy as it was in the one of Karam and Schraefel (2005). Bernsen and Dybkjær (2009, p. 77) indeed did not consider these terms in physical aspects "*but, rather, as what the recipient of a representation can perceptually inspects for as long as it takes … versus items that change beyond the control of the recipient*". Therefore, in their taxonomy, a static representation corresponds to the situation where we must act to change it (for instance, the alarm clock on your phone that does not stop

without intervention), while a dynamic representation does not need any intervention from the recipient of the representation to change as it does on its own (for example, when your telephone rings). Therefore, the term "static" "*includes all that is static in the ordinary sense plus short-term cyclic change which also allows freedom of perceptual inspection*" (Bernsen & Dybkjær, 2009, p. 77).

In short, the modalities included in the generic level have been based on five distinctions which make "*fundamental differences to usability in user–system interaction*" (Bernsen & Dybkjær, 2009, p. 79): linguistic vs. non-linguistic; analogue vs. non analogue; static vs. dynamic; arbitrary vs. non-arbitrary; and graphics vs. acoustics vs. haptics.

Now that the basics have been described, here is a small lexicon of more precise terms used at the atomic and subatomic levels given by Bernsen & Dybkjær (2009, pp. 81-83):

Label/Keyword: a meaningful expression, such as a word or phrase, which is used independently in any type of language.

Notation: "*either a formal language or something more loosely defined, such as a somehow organised or restricted set of expressions developed for some purpose*" (Bernsen & Dybkjær, 2009, p. 81).

Image: perceptual appearance of something in any spatial dimension, in time or out of time, in a superficial or realistic way.

Map: not considered as a compositional diagram, even though it is a special type, because map is a too familiar concept to be considered only at the sub-atomic level.

Compositional diagram: image that in some way takes on the perceptual appearance of something else ("analytical image).

Conceptual diagram: representation, thanks to spatial and temporal means, of relationships considered as conceptual.

Graph: representation, thanks to spatial and temporal means, of quantitative data.

Obviously, the taxonomy developed by Bernsen and Dybkjær (2009) is much more complex than the summary above, but the background provided makes it easier to understand what the main abbreviations in this table correspond to, although it is in the legend. Furthermore, although the representation of gestures can be described by Bernsen and Dybkjær's taxonomy, it also gives an idea of how to represent the information provided by the device with which users communicate.

Super level	Generic level	Atomic level	Sub-atomic level
Linguistic modalities →	1. Sta. an. graphic ele.		
	2. Sta.-dyn. an. acoustic ele.		
	3. Sta.-dyn. an. haptic ele.		
	4. Dyn. an. graphic →	4a. Sta.-dyn. gest. disc.	
		4b. Sta.-dyn. gest. lab.	
		4c. Sta.-dyn. gest. not.	
	5. Sta. non-an. graphic →	5a. Written text →	5a1. Typed text
			5a2. Hand-writ text
		5b. Written lab. →	5b1. Typed lab.
			5b2. Hand-writ lab.
		5c. Written not. →	5c1. Typed not.
			5c2. Hand-writ not.
	6. Sta.-dyn. non-an. acoustic →	6a. Spoken disc.	
		6b. Spoken lab.	
		6c. Spoken not.	
	7. Sta.-dyn. non-an. haptic →	7a. Haptic text	
		7b. Haptic lab.	
		7c. Haptic not.	
	8. Dyn. non-an. graphic →	8a. Dyn. written text	
		8b. Dyn. written lab.	
		8c. Dyn. written not.	
		8d. Sta.-dyn. spoken disc.	
		8e. Sta.-dyn. spoken lab.	
		8 f. Sta.-dyn. spoken not.	
Analogue modalities →	9. Static graphic →	9a. Images	
		9b. Maps	
		9c. Compos. diagrams	
		9d. Graphs	
		9e. Conceptual diagrams	

Legend
an. = *analogue*
compos. = *compositional*
disc. = *discourse*
dyn. = *dynamic*
ele. = *elements*
gest. = *gesture*
lab. = *labels /keywords*
non-an. = *non-analogue*
not. = *notation*
sta. = *static*
writ. = *written*

Figure 6- A taxonomy for gesture information representation (Bernsen & Dybkjær, 2009, p.73)

Super level	Generic level	Atomic level	Sub-atomic level
	10. Sta.-dyn. acoustic →	10a. Images	
		10b. Maps	
		10c. Compos. diagrams	
		10d. Graphs	
		10e. Conceptual diagrams	
	11. Sta.-dyn. haptic →	11a. Images →	11a1. Gesture
		11b. Maps	11a2. Body action
		11c. Compos. diagrams	
		11d. Graphs	
		11e. Conceptual diagrams	
	12. Dyn. graphic →	12a. Images →	12a1. Facial expression
		12b. Maps	12a2. Gesture
		12c. Compos. diagrams	12a3. Body action
		12d. Graphs	
Arbitrary modalities →	13. Sta. graphic	12e. Conceptual diagrams	
	14. Sta.-dyn. acoustic		
	15. Sta.-dyn. haptic		
	16. Dyn. graphic		
Explicit structure modalities →	17. Sta. graphic		
	18. Sta.-dyn. acoustic		
	19. Sta.-dyn. haptic		
	20. Dyn. graphic		

Figure 7- A taxonomy for gesture information representation (Continued) (Bernsen & Dybkjær, 2009, p.74)

3. GES REVIEW IN CHOSEN APPLICATION DOMAINS

In this part, I explain how I reviewed the forty-two chosen GES as well as how I classified the gestures acquired in them. Moreover, I propose some adjustments in these classifications in order to facilitate the classification of the gestures by the managers.

3.1 Gesture set selection

For this study, I selected 42 GES related to the five chosen applications domains, that are as a reminder: Medical Systems and Assistive Technologies; Crisis Management and Disaster Relief; Entertainment; Human-Robot Interaction; and TVs. From these 42 studies, 4 were realised in the Medical systems and assistive technologies' domain, 9 can be used for crisis management and disaster relief, 17 in Entertainment, 6 in human-robot interaction, and 11 in the development of smart TVs. Among them, 4 studies could be used in two or more domains, like it is the case for Chen et al's (2017) study that aims to extend gestural interaction from hands to other body parts, and that therefore could be related mostly to crisis management, entertainment, and TV.

The choice of these studies was based on three main criteria. The first was of course that the gestures acquired were only through the use of a camera (camera, webcam, Kinect) and that the utility of the set of gestures produced was to be used to communicate with an interface allowing the collection of 3D gestures. The second was that a set of concrete gestures could be proposed as a result of the study and that these were well represented either in written form or visually through diagrams or images. The third was that the studies should only consider gestures made without wearing sensors, without touching the communication interfaces and without talking to the interfaces. In other words, the studies should only look at mid-air gestures.

Regarding the first criterion listed, the set of GES is divided as follows: 3 studies were conducted using webcams, 24 using at least one camera (professional or not), and 15 using Kinect. In addition, although Leap motion is described by Ultraleap (2022) as the most reliable hand gesture detection technology with unparalleled quality, I have chosen not to include this technology in my gesture review because it still requires you to be at a maximum of 80 centimetres away, and preferably between 10 and 60 centimetres (Ultraleap, 2022). This

distance constraint was therefore rather problematic in view of the chosen domains. Indeed, if the surgeon operating has to move away from the patient to get closer to the interface in order to view certain things, this would not be optimal for example. Another example would be in the field of entertainment or TV, and especially in the case of games which require you to be quite active, it would not be practical to have to stand so close to interact.

Then, for the second criterion, I only took into consideration the gestures that had been proposed most often by the participants of the different studies. However, I sometimes had to sort through the gestures listed because the researchers had simply put all the different rates of agreement with all the proposed gestures or had listed the proposed gestures with the frequency with which this gesture was returned. Indeed, as in the case of the elicitation study by Wu et al. (2014), a table (see Appendix 1) was provided with the frequency of occurrence of the different gestures for each referent, sometimes as many as 11 different gestures proposed for one referent. Thus, when this happened, I selected the most proposed gesture of the list when there was a large difference between the first and the second (i.e., in Wu et al's study (2014), to increase the volume of TV, 21 of 24 participants agreed on a "*swipe up*" gesture), and I took two gestures in the case where there was a little difference between both (i.e., in Wu et al's study (2014), 5 participants agreed to "spread their hand" to open the menu while 5 others agreed to "snap their fingers"). Moreover, as explained by Bostan et al. (2017, p. 1131), who proposed a set of on-skin gestures, users tend to limit their mental and physical load when interacting, and therefore tend to propose usual gestures when they are involved in GES. This phenomenon is called the legacy bias and cannot be avoided. In addition, Bostan et al. (2017, p. 1131) add that more than producing habitual and intuitive gestures, study's participants tend to think that these gestures better fit the intended actions than do exclusive gestures. In their intuitive gesture sets, the problem was that most of the gestures were "in-context" and that one gesture was used for several actions that could happen in the same context, which is not possible. Thus, Bostan et al. (2017, p. 1131) considered that the exclusive gesture set (see Appendix 2) they created was a better alternative although it possesses lower agreement rates, and that is the one I used in my analysis. Furthermore, some studies, like the one of Jurewicz et al. (2018), asked two different types of users to take part in their study, in their case novices (undergraduate and graduate students) and experts in the anaesthesia domains. In this case, I considered all the proposed gestures.

Besides, in order to have a clear view on the representation of the gestures, I have classified all the articles I had selected thanks to the way of representing the gestures described

in Bernsen and Dybkjær (2009) taxonomy. This allowed me to sort out all the pre-selected GES more easily in order to only keep those that were well represented with either analogue or linguistic modalities. Among the 42 GES selected, solely 7 were described uniquely by linguistic modalities, and more specifically static non-analogue graphic modalities, and even more specifically static written typed text or static written typed label. The remaining 32 GES were a mix of analogue and linguistic modalities, sometimes with written typed text, and sometimes with static written typed label.

Finally, regarding the third criterion of limiting myself to studies that only includes mid-air gestures, I have always followed this but made an exception for the Felberbaum and Lanir (2016) study which was based on gesture interaction with the feet and which used, in addition to the camera, a bounded area on the floor that the user could not go beyond. The reason I made this exception is that I really wanted to try and evaluate the 3D gestures performed by all parts of the body, from head to toe.

In addition, in order to have a clearer view of all the different gestures performed, when two or more identical gestures (e.g., the thumbs-up sign) were used for several referents in the same GES and had the same characteristics regarding the taxonomy used (see section 3.2), they were considered as one gesture in my table. But if one of the dimensions of the taxonomy was different, they were considered as different gestures. Furthermore, if two gestures were identical in two different GES, they were considered as two different gestures.

Considering all this, I ended up with a set of 571 gestures whose 88 were used for two or more tasks and therefore could be used to perform a total of 696 referents. Below is a table of the most common types of gestures in all areas:

Gesture Type	**Frequency**
Swipe	98
Towards the right	81
Upwards	77
Towards the left	68
Point	62
Downwards	52

Circular	33
Forward	24
Push	22
Open hand/arms	21

Table 2- Most common gesture types

As you can see in the table above, the type of gestures that was the most performed by the participants in this review was the *Swipe* move with 98 occurrences. Among those swipe moves, 22 were directed to the left, 21 down, 20 to the left and 17 up, the last 18 were swipes with no specific direction imposed. The second type of gestures the most used was gestures that must be directed to the right, or that used the right part of the body, whether it be hand, arm, leg, or foot, with 81 occurrences. By removing the 18 swipe moves, there are still 63 moves involving a direction to the right or the right side of the body. At the bottom of the podium, we find the movement to direct upwards with 77 occurrences. Less the 17 swipe moves, it is still 60 moves that implies upward direction, whether it be with thumb (thumb-up), or by pointing with hands or arms.

Now that the most used gestures have been listed, let's analyse what were their most common purposes in the different application domains:

- *Swipe left* gestures were used most often in crisis management to move an object (generally an image or video) to the left but still keeping it on the screen. In entertainment and with TVs, these gestures were mainly used to say "next" video, music, or channel. Concerning operating rooms, this type of gesture was mainly used to pan the camera or move the image to the left. Finally, with robots, this gesture was mainly used to indicate the robot to turn left.
- *Swipe right* gestures were mainly used to make the exact opposite of what Swipe left gestures aimed to achieve. Thus, for example, in entertainment and TV, these were used to say "previous" video or channel instead of "next".
- *Swipe up* gestures for their part were mainly used in TV and entertainment domains to increase the volume; in operating rooms for panning upwards or moving the image up, but also to increase flow of anaesthesia gas (according to novices in the domain); and with robot to make them go higher (drones) or to make them move backward or even to increase their speed.
- *Swipe down* gestures were mostly the exact opposite of swipe up gestures.

- *Thumb-up* signs were mainly used to say “OK”, “Yes”, or “Accept” whatever the domains.
- *Point* gestures were mainly used to select items in TV, crisis management, and Entertainment, or indicate a specific direction. Which does not differ from their definitions in Aigner et al.’s (2012) study.
- *Circular* gestures, whether it be by drawing a circle, or moving hands or fingers clockwise or counter-clockwise, were mainly performed in all domains to represent rotation (clockwise = right rotation and counter-clockwise = left). Nevertheless, in Brito et al.'s (2019) study in human-robot interaction, rotate clockwise was used to ask the robot to “redo the task”, which clearly looks like the arrow that refreshes a web page.
- *Forward* movements generally include *Push* gestures with hands or fingers, which were mostly performed to select an item with large displays in crisis management, to open a menu in TV applications, or to indicate to a robot to move forward. Moreover, forward movements were also performed in entertainment and robot-interaction with foot in order to indicate the robot or the avatar (in games) to walk forward.
- Finally, *Open hands/arms* gestures (gestures that move smoothly from a closed to an open position) were mainly performed in Entertainment and TV to undo tasks that have already been completed, such as closing your fist or arms to mute a TV or pause a video. Thus, opening hands or arms was used to mean “unmute” or “play” in the examples.

Furthermore, from my observations, I realised that some referents (not mentioned above) appeared in all the selected application domains but for which the gestures were not the same because they depend on the size of the system with which the users interact. The most common one is *Zoom (in/out)*. Zooming in was always performed by spreading fingers or hands whatever the domains. Sometimes with a pinch handshape, sometimes simply by distancing hands without considering a specific hand shape. On the opposite, zooming out was performed by bringing together hands or fingers. These tasks are present 38 times among all referents.

3.2 Taxonomy used to classify the gesture set

To classify the gestures in the most optimal way in my opinion, I chose to combine both the classification of Aigner et al. (2012) with that of Vafaei (2013). Indeed, I found that the way of analysing the nature of gestures in Aigner et al. was better than that of Vafaei, the latter

having added the notion of abstract gestures, a class which did bring too much subjectivity when assessing gestures that come from GES conducted by other researchers.

Moreover, since Vafaei (2013) also used the notion of context to describe the gesture, I felt that there was a kind of duplication of effort because generally when users perform abstract gestures, they use them several times to perform different tasks (which makes them "In-context"), the users who use abstract gestures the most being those who do not know the range of gestures they can use (Vafaei, 2013, pp. 18-19).

In addition, I have not considered the notion of range of motion either, even though I must admit that it has its interest when we conduct our own study of gesture elicitation. This is why this notion will also be present in the taxonomy proposed later in section 3.5. In fact, since I was doing a review and only had access to static analogous modalities in the form of images as well as static linguistic modalities in the form of text, it was impossible for me to have precise information on the range of motion of gestures.

Furthermore, I did not consider that the dimensionality characteristic of gestures from Vafaei (2013) taxonomy was really useful given that the aim of this review of studies was to collect and analyse gestures performed in a 3D space. Furthermore, according to me, users have no idea of dimensionality when they perform their gestures. At present, people tend to prefer simplicity, so they certainly do not consider the axis in which they perform their gestures.

Taking all this into consideration, below is the taxonomy used to analyse the 571 gestures:

Gesture mapping	
Nature	● Pointing ● Semaphoric ● Pantomimic ● Iconic ● Manipulative
Form	● Static ● Dynamic ● Stroke
Binding	● Object-centric ● World-Dependent ● World-independent
Temporal	● Discrete ● Continuous

Context	• In-context • No-context
Gesture characteristics	
Complexity	• Simple • Compound
Body Part	• Head • Shoulders • Hand • Arm • Foot
Handedness	• Dominant • Non-Dominant • Bi-manual
Handshape	• Flat • Open • Bent • Curved • Index Finger • Fist • Free • ASL Shapes • …

Table 3- Used taxonomy for Reviewing

3.2.1 Classification hypotheses

In order to classify all the gestures, I was asked to make a number of assumptions. Indeed, not all the information taken into account was specifically given, such as handedness. Consequently, here is a description of the several assumptions made:

➢ ***H1***: *Handedness: If hand/arm not imposed → Dominant*

When the hand or arm to be used was not imposed, I considered that people would automatically use their *dominant* hand or arm by default, as they would in everyday life (a right-handed person rarely uses their left hand to use actions, and vice versa).

➢ ***H2****: Handedness: If hand/arm imposed → Non-dominant*

Conversely, if the arm or hand to be used was imposed, I automatically considered the gesture to be non-dominant because the choice is therefore not left to the user.

➢ ***H3****: Handshape → On the basis of the image of the gesture if possible*

When the written description of the gesture did not specify the shape of the hand to be used to perform the gesture, I relied on the picture that represented the gesture. If it was possible to guess from the picture what shape was used, then it was used. For example, in the study by Wu et al (2014), gestures including a swipe were clearly represented with a flat hand.

- ***H4***: *Handshape → Gestures without a clear description of the hand shape or not clearly visible in the image → Free hand*

On the contrary, when the image did not allow the shape of the hand to be guessed, the gestures were automatically considered with a free hand. Furthermore, when the gestures took into account only the arm and not the hand, they were also automatically considered with a free hand because no matter what the shape of the hand, they would be understood by the system.

- ***H5***: *Handshape → When taping two or more fingers together or with the palm → Joined fingers*

In the GES by Chan et al (2016), which studied single-handed microgestures, several gestures involved tapping two or more fingers together or tapping several fingers with the palm, which did not correspond to any hand shape. However, since they involved joining a few fingers together, I considered that these gestures corresponded to a "joined fingers" hand shape.

3.3 Gesture set analysis

After explaining the used taxonomy, as well as describing the different assumptions I made, the different dimensions of the taxonomy can be analysed more deeply.

3.3.1 Nature

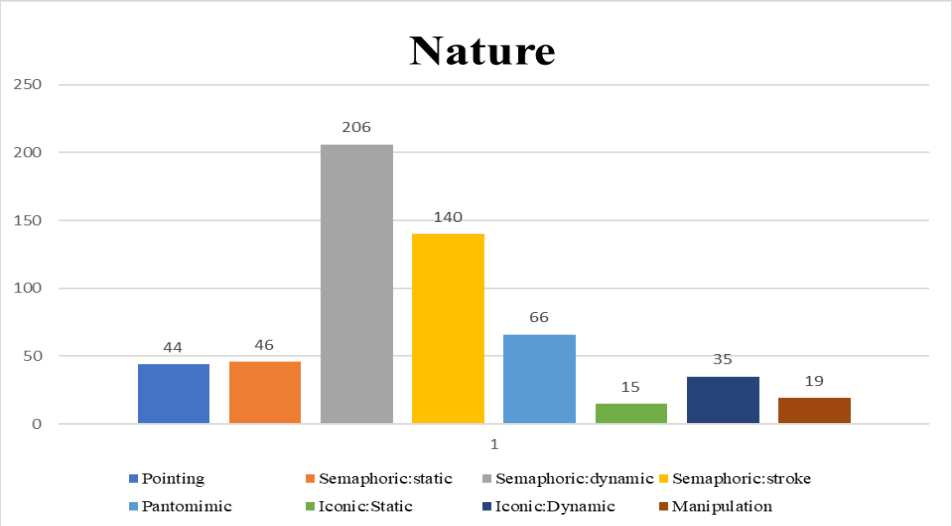


Figure 8- Nature distribution of the 571 gestures

Figure 8 above represents the breakdown of nature dimension across the 571 gestures. As you can see, semaphore gestures make up the bulk of this set of gestures, with 392 gestures, of which 346 are dynamic. However, while the most common gestures were swipe gestures (98 out of 571, i.e., almost 20%), dynamic semaphoric gestures still outnumber stroke gestures. The third most common type of gesture found in this set of gestures was pantomimic gestures. These were most often found in the field of entertainment, and a bit in interaction with robots, as users tend to imitate in games the actions they wish to perform when possible, or the actions they want the robot to perform.

3.3.2 Form

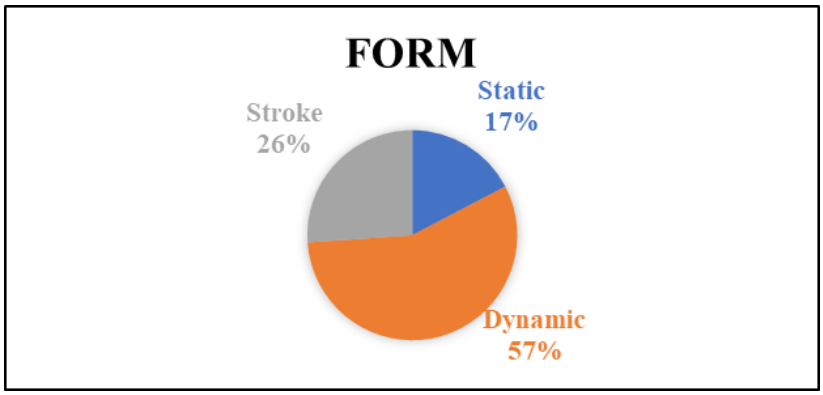


Figure 9- Form distribution

Regarding the distribution of the form of gestures, I found that although dynamic gestures were the norm in HCI, with 57% dynamic gestures and 26% stroke-like gestures, there were still 17% static gestures performed in 3D spaces. This therefore helps to justify why I did not simply use the taxonomy developed by Hiyadi (2016) for the purpose of analysing 3D gesture capture in human-system interactions. The static gestures that came up the most across the different ones were of course the pointing gestures generally used to select item or Turn On/Off TV, but also “fist”, the "Thumb-up" and the "Thumb-down" signs, as well as the hand or hands flat in front of participants to signal the system, in general, to stop what it is doing.

3.3.3 Binding

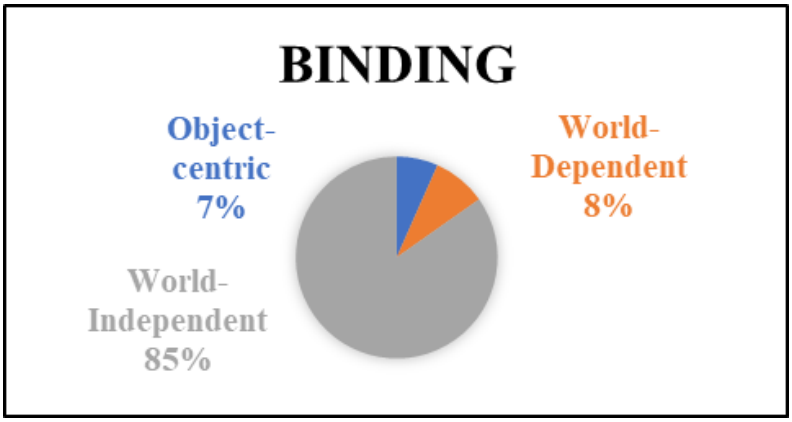


Figure 10-Binding distribution

As I investigated mid-air gestures, it is not surprising that most of these gestures were world-independent in the sense that gestures could be performed wherever users wanted to execute them in the air. World-dependent and Object centric gestures mainly occurred in crisis management applications, where they have to manipulate large amounts of various items, but also in medical systems when browsing large quantities of images.

3.3.4 Temporal

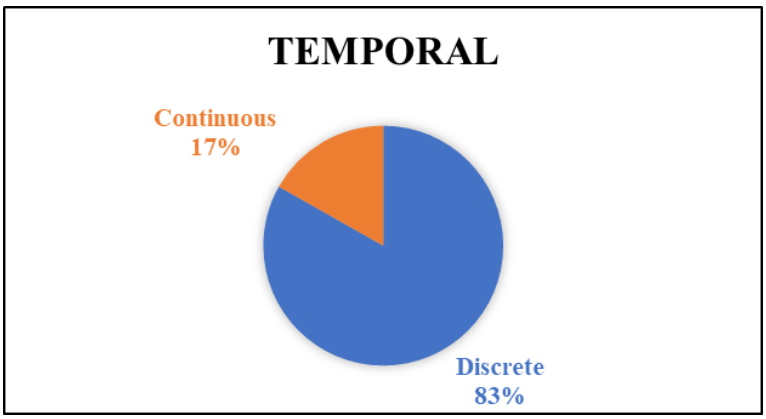


Figure 11- Temporal distribution

In terms of temporal allocation, I make the same point as Vafaei (2013, p. 41) about manipulative gestures. Indeed, gestures involving grabbing or panning are always continuous. It is therefore not abnormal to see the percentage of continuous gestures (17%) being very close to the combined percentages of gestures considered object-centric and world-dependent (15%). The slight difference between these percentages can be explained by the fact that in the medical field, sometimes the zooming gestures are continuous; and in the TV field, sometimes the users want to continuously increase or decrease the sound, so they need to know where they are while performing the gesture, which makes it continuous.

3.3.5 Context

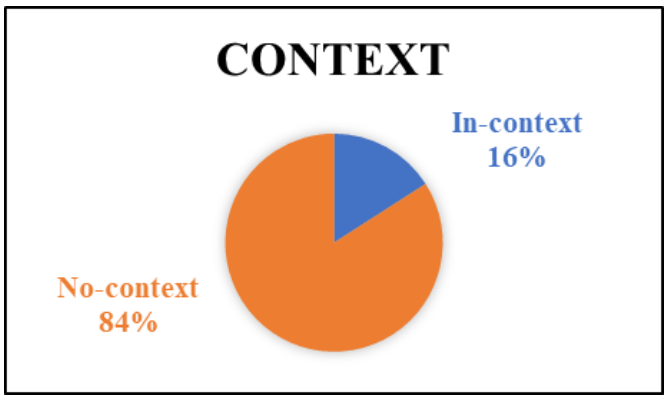


Figure 12- Context distribution

As shown by figure 12, gestures "in context" are clearly not the norm, and fortunately so. This type of gesture is mainly found in crisis management and medical applications, where users need to process large amounts of data without wasting too much time remembering the gesture they need to make to perform a specific task. Moreover, as I reviewed one GES about micro-gestures on only one hand (Chan et al ,2016), one about on-skin gestures performed with two hands (Bostan et al., 2017), this increased the number of gestures "in-context" because with such a constraint as being limited to one or both hands, it is inevitable that participants do not revert to gestures already used. Nevertheless, while Bostan et al. (2017) proposed a set of exclusive gestures that could only be used in different contexts, as a counterproposal to the intuitive gestures found during their studies (which were problematic because they could be used twice in the same context); I was surprised by the diversity of gestures proposed by the participants who were only able to use one hand in the study made by Chan et al. (2016).

3.3.6 Complexity

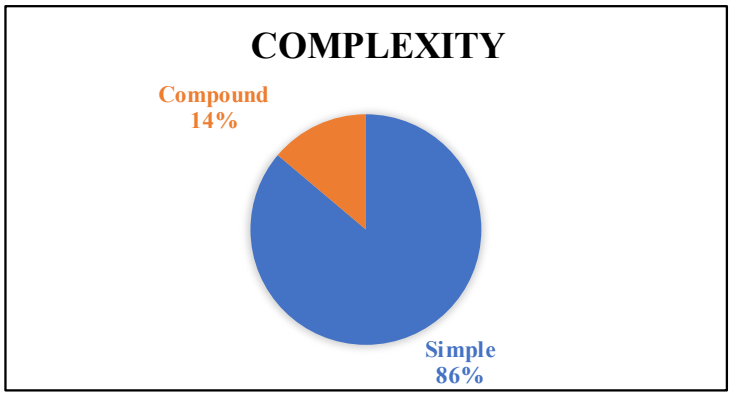


Figure 13- Complexity distribution

Whatever the domains of application, participants tended to propose simple gestures to perform tasks, one of the reasons being that this allows them to easily remember what the

gesture is. Compound gestures, on the other hand, appeared mainly in the GES conducted by Madapana et al. (2019) on "*Obtaining gestural guidelines for PACS in neurosurgery*", which required users to use only both hands. Typically, the gestures took place in two phases: the first was to point the hand in a particular direction, and the second to move it in a certain direction. The rest of the compound gestures have mainly appeared in the TV area with repetitive swipes to increase or decrease the sound, as well as changing several channels at once.

3.3.7 Body part

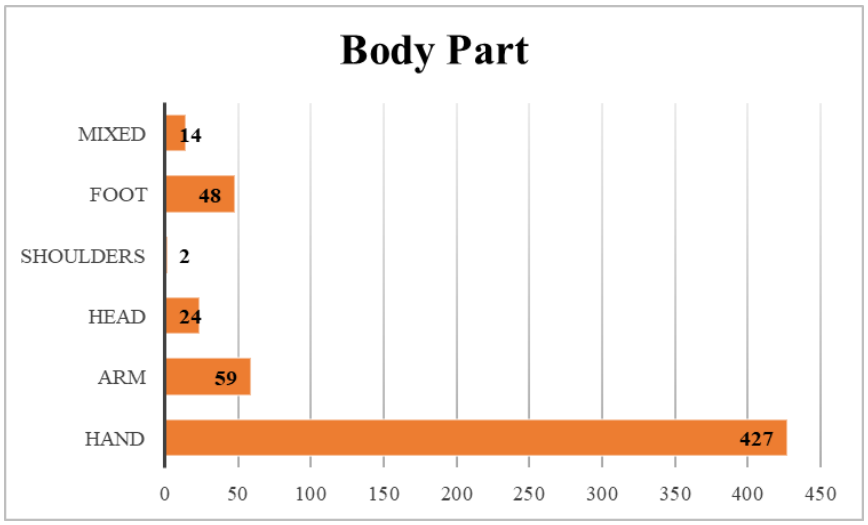


Figure 14- Body part distribution among 571 gestures

Even though users have the possibility to use their whole body, they have a strong tendency to always use their hands as shown in Figure 14, with hands used in 427 of the 571 gestures analysed. This proves the interest of using notions such as handedness, as well as the shape of the hand to differentiate between gestures. Given the areas analysed, this is not surprising. In the medical field, doctors clearly do not have time to use anything other than their hands and arms. As for TVs, users are often in a seated position, so they don't really know how to use anything other than their upper body. For crisis management, since data management has to be fast, participants also tend to use mainly hand gestures, as these are the most natural and intuitive. In the areas of entertainment, and human-robot interaction, users also tend to use their feet more often for moving actions but will also use their arms and hands for the majority of their actions.

3.3.8 Handedness

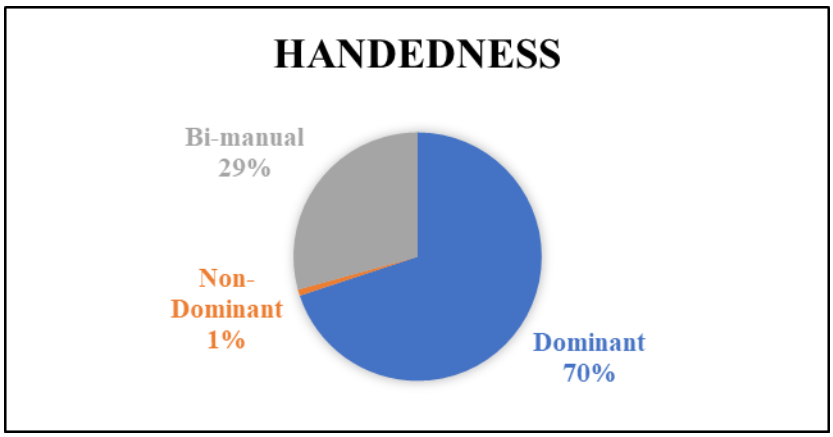


Figure 15- Handedness distribution

Figure 15 above represents the handedness breakdown between the 571 gestures. As shown, most of the gestures were dominant, which was partly influenced by assumptions H1 and H2 described above. In general, when the non-dominant hand or arm was involved in gestures, it was used as a kind of reference, and accompanied by the dominant hand moving to communicate with the system, making these types of gestures bimanual. Moreover, bimanual gestures were usually performed to differentiate from dominant swipe gestures, like for "Pan" tasks in medical systems, or "Zoom In/Out" when operating on large displays.

3.3.9 Handshape

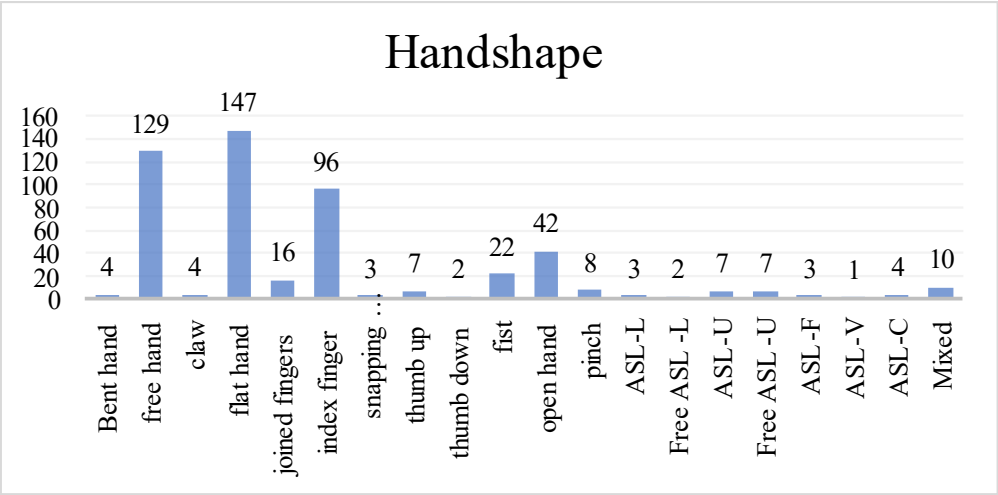


Figure 16- Hand-Shape distribution

Finally, figure 16 provides a clear view of the hand shape allocation from the hand and arm gestures. "Flat hand" is the shape that was the most used, principally to perform "swipe" gestures as well as "Pan" gestures, it was also used to perform "Stop" signs and sometimes to

represent zooming gestures on large displays. The second shape that occurred mostly is “free hand”, but this is not really representative as it is due to assumption H4. Afterwards, we found the “index finger” shape, which was principally used for pointing gestures, but also for “swipe” ones. Furthermore, the fact that ASL gestures are not more present in the 42 studies analysed is probably caused by the lack of knowledge of this language by most users, who therefore do not think specifically about using these gestures. The distribution would probably be very different in a study of deaf or hard of hearing people with full knowledge of this language. In addition, regarding the 10 “mixed” handshapes, it was generally movement involving coming from a “fist” shape to an “open hand” shape, and vice versa, to perform “Play/Pause” or “Mute/Unmute” actions in Entertainment and TV domains.

3.4 Main issues when classifying gestures

Although I had made some assumptions to simplify the classification of the gestures, I encountered some problems in classifying this set of 571 gestures. The following are the main problems noted:

- ***P1***: When manipulative gestures were used to perform “grab” actions, sometimes this was represented with a “claw” handshape, and sometimes with a close fist. Consequently, I was confused when I only had a linguistic written modalities to describe a “grab” action because I did not know if I should consider it as “claw” or a “fist” shape. A possible solution could be to create a “grab” handshape and add it to the set described in Table 1.
- ***P2***: Sometimes, specifically in entertainment applications, but it could also apply to human-robot interactions, the gesture performed included other parts of the body that the ones mentioned in Table 3. As a matter of fact, of all the gestures, 9 were performed using the torso, 6 using the legs, and 10 using the whole-body posture. And even if this represents few gestures, they were more present than the gestures made with the shoulders (only 2 occurrences), which are taken into consideration. Therefore, I suggest that these parts could be optionally added to the taxonomy depending on the intended application areas.

3.5 A taxonomy for classifying 3D-acquired gestures

Following the analysis of these problems, as well as the use of the taxonomy above (table 3) to classify the 571 gestures from the 42 GES, here is the taxonomy that I propose to classify the gestures acquired by 3D camera:

Gesture mapping	
Nature	• Pointing • Semaphoric • Pantomimic • Iconic • Manipulative • Emblematic
Form	• Static • Dynamic • Stroke
Binding	• Object-centric • World-Dependent • World-independent
Temporal	• Discrete • Continuous
Context	• In-context • No-context
Gesture characteristics	
Complexity	• Simple • Compound
Body Part	• Head • Shoulders • Hand • Arm • Foot • Leg • Torso • Whole
Handedness	• Dominant • Non-Dominant • Bi-manual

Handshape	• Grab • Flat • Open • Index Finger • Fist • Free • ASL Shapes • …
Range of Motion	• Small • Medium • Large

Table 4- Proposed taxonomy for classifying 3D-acquired gestures

In blue, in Table 4, are all the notions that I did not use to classify the gestures but that seem useful to use when conducting its own GES. All other dimensions used are to be understood as previously defined.

I did not consider the classes developed by Hiyadi (2016) because I did not have any studies that included facial expressions and I wanted to mainly focus on gestures, even if facial expressions could be detected by 3D communication interfaces. Nevertheless, I consider that the class relating to the **emblematic** gestures it proposes is important in the sense that it is strongly linked to culture. Therefore, it seems important to me to identify such gestures, as they are culturally related and should therefore probably be adapted in the case of an internationalisation of the device.

Furthermore, as explained above, in order to solve the P1 problem, I suggest having a specific handshape that would represent a "grabbing" shape. Another solution could be to clearly define a handshape for "grasping" movements between "claw" and "fist» but adding a new shape could increase the range of gesture possibilities.

Regarding the addition of body parts, this is mainly a necessity in human-robot interactions as well as in the entertainment domain, but the addition of body parts to be detected in the end also allows both developers and participants to increase the likelihood that they use other body parts than the one mentioned by Vafaei (2013).

Finally, range of motion is an important dimension to consider, as it clearly allows managers to differentiate between the same types of gestures, such as two "slide left" gestures both performed with the hands, but with a different range of motion for example.

4. MANAGERIAL IMPLICATIONS

In this section, I depict the managerial implications that are linked with the development of 3D devices as well as the implications behind the taxonomy proposed in section 3.5. In addition, I have conducted a small online survey to gather the views of potential end-users on certain aspects, to find out what they consider necessary or not. However, this survey can only provide exploratory data as only 31 participants took part.

4.1 Exploratory survey analysis

Thanks to this online survey (see Appendix 3), I could gather some interesting exploratory data. The distribution of the participants who took part in the survey is as follows: with regard to gender, there were 22 women and 9 men; with regard to age, there were 7 people under 24 years of age, 23 people between 24 and 65 years of age, and 1 person over 65 years of age; with regard to occupations, 12 participants were students, and the remainder were variously distributed, ranging from teachers to accountants to pharmacy assistants. Furthermore, what is particularly interesting from a managerial point of view is that only 8 participants were familiar with such technologies (Alexa, Google Assistant, Kinect and Leap motion). Therefore, the remaining 23 represent potential future users.

When the participants were asked the question "In which places (public or private) can you imagine using this kind of technology?", the answers confirmed the importance of the application areas I had chosen. Most of the answers indeed include home, shops, hospital and offices. In addition, 21 participants support the emergence of these types of communication interfaces in their current environment, the reason being that they considered that it was more fun (gamification), that it could save time, and that it could also make daily life easier. However, several of them also mentioned that they did not want these technologies to completely replace humans. As for the 10 people who were against, they surprisingly included 2 people under the age of 24, an age group I would have thought was 100% in favour, and not surprisingly, the person over 65. The main reasons mentioned against the appearance of these technologies in people's immediate environment were dehumanisation in the sense of reducing human-to-human contact, lack of vision as to the usefulness it could bring them, as well as the fear of not knowing how to control the interface properly.

Regarding having to use simple or compound gestures with the system, while 15 participants said they preferred simple gestures, 14 considered that it did not matter to them. The last two participants preferred the idea of using compound gestures with the aim of the second gesture being a validation of the first. Regarding the fact that gestures could be used in various context (when they are "in-context"), 12 participants said they were bothered by having to use the same gesture for several actions, mainly because they were afraid that the system would misunderstand the task they wanted to perform, or that they would forget what the gesture was for. The rest of the participants, on the other hand, did not see this as a problem because they did not find it complicated, and even thought it was a good thing if the system could understand the same gesture in different contexts.

Furthermore, I also investigated the time (in seconds) users consider the computer should take to provide its response. For the simple tasks (turning on the TV, turning up the volume, turning on the light, etc.) the average response time of the participants is 5.37 seconds, with a standard deviation of 7.53 seconds. The lowest response given was 1 second, and the highest 30. Afterwards, for average tasks (simple sorting (alphabetical order) in a file, downloading a small file, etc.), the average response time of the participants is 13.13 seconds, with a standard deviation of 13.9 seconds. The lowest response given was 1 second, and the highest 60. Finally, for complex tasks (asking to perform a complex calculation, downloading a large file, etc.), the average response time of the participants is 805 seconds, with a standard deviation of 805,85 seconds. The lowest response given was 1 second, and the highest 18000. With regard to the different standard deviations, the averages for simple and average tasks can be considered, but not the average for complex tasks because of its high value.

Afterwards, I analysed how users would like to see the gestures represented if they had the opportunity to have a learning guide to learn the pre-recorded gestures in the system. I have therefore used the criteria of McAweeney et al (2018) to assess this. Participants were given the choice between a simply written representation (1st person), or a visual representation (3rd person), and it was found that 27 of the 31 participants preferred a visual representation. Subsequently, they had to choose between a mirror, bird's eye or side view. The sample showed that 16 participants preferred a mirror view, while 14 preferred a side view. Only one person chose a bird's eye view.

Then, I built on how Rose et al (2012) had used an empirically validated scale to analyse some of the data, in order to analyse the following criteria: users' skills; their confidence in the system; their level of satisfaction (for those who had already been able to use such technologies); the control they feel they have over the interface; and customisation.

In terms of the *skills* that users considered they needed to interact with the system, 17 of the respondents considered that they need a quick learning guide to learn how to interact with the interface, 10 considered that they can manage with their knowledge, 3 considered that they have the necessary skills, and one said that he was more competent than the others (the only respondent who mentioned the jumping motion technology). For this reason, section 4.4 will mention how to represent gestures in a quick learning guide.

For *trust* in the system, 16 respondents considered this type of system to be merely reliable, 13 considered the quality of the provider to be sufficient information to trust this type of system, and only 2 considered that they could trust this type of system completely without any uncertainty.

Concerning the level of *satisfaction* among people familiar with this kind of system, out of the 8 people, 5 were satisfied and still use the system now (or in the near future) and 3 were satisfied but no longer use this kind of technology.

As for the level of *control*, among the regulars, the answers varied between those who had a correct control but still made a small error from time to time, and those who felt they had a fairly good control of the system but still made some errors.

Finally, the *customisation* of the system remains an important notion according to the results of this survey. Indeed, only 3 of the 31 respondents were satisfied with the pre-recorded commands; 9 were generally satisfied but still wanted to be able to adapt some gestures; 12 wanted to be able to configure themselves the gestures to be used to communicate with the interface but wanted these gestures to be proposed and that they simply had to choose; and finally 7 wanted to be able to configure the gestures themselves from scratch.

In addition, in view of the different requirements listed above, I also analysed the body parts they wanted to use, as well as the importance they gave to the following requirements: ease of execution of the gestures; memorability of the gestures; comfort; time response; the fact of being able to come "as you are"; as well as the importance of using a vocabulary of natural and intuitive gestures.

Concerning the *parts of the body* that people wanted to use in order to perform the gestures, 20 of the 31 respondents wanted to use only their hands and arms; 6 wanted to be able to use their whole body; 4 wanted to be able to use the whole upper part of their body; the last one wanted, in addition to using their body, to use their hands and arms.

Needs ranging from ease of execution to using an intuitive vocabulary were rated by participants on a scale of 1 (useless) to 7 (indispensable). The table below shows the average responses for each of the requirements:

Requirement	Average score
Ease of execution	6.16
Memorability	6.19
Comfort	5.81
Responsiveness	5.87
“Come as you are”	4.94
Intuitiveness	5.9

Table 5- Average score of the different requirements

The average scores above show that the two most important things for respondents to my questionnaire are ease of execution and memorability of the gestures, followed by using intuitive gestures and the system's speed of response and comfort in executing the gestures. Not having to carry a device is less important to them, but not useless.

4.2 Challenges behind interfaces requirements

In addition to defining the basic requirements of communication interfaces developed for human-machine interaction, Wachs et al. (2011, pp. 62-64) also explained the challenges related to these requirements.

4.2.1 Price

Depending on the application domains targeted, the hardware needs will differ. According to Wachs et al. (2011 p. 62), the main challenge for a developer is to identify the hardware necessary for the system to run smoothly regarding the available development budget. Therefore, managers should work together with the developers, to ensure that sufficient budget is allocated regarding the necessary hardware.

4.2.2 Responsiveness

The challenge in terms of responsiveness is to find the right and simple characteristics on the computational side (Wachs et al., 2011, p. 62). In addition, Nielsen (1993 and 2010) describes the 3 important limitations in terms of response times. A user will consider that the system reacts instantaneously if the system response occurs maximum **0.1 second** after the request. This type of responsiveness is needed for direct manipulation. Then, if the system takes **1 second** to respond, Nielsen (1933 and 2010) explain that users will notice the delay but will not feel like they are not controlling the system as it "*keeps the user's flow of thought seamless*". This responsiveness level is necessary to ensure a good navigation through the system. Finally, the last limit to keep the user's attention on the dialogue is **10 seconds**. Nielsen (1993 and 2010) explains that beyond this period, feedback should be provided on the progress of the current task, while allowing the user to continue with other tasks.

4.2.3 User adaptability and feedback

As shown by my exploratory survey, users tend to want a system that is adaptable to what they prefer, rather than a system that does not possess real-time processing and in which the core algorithm was trained offline (Wachs et al., 2011, p.62). Thus, the challenge for managers is to implement this real-time flexible algorithm. Moreover, an additional problem depicted by Wachs et al. (2011, p.62) is that presenting feedback should not add cognitive load.

4.2.4 Learnability

According to Wachs et al. (2011, p. 62), the learning rate depends on three aspects: the task in itself; the experience that the user possesses with HCI interfaces; as well as the cognitive competences of the user. Therefore, this makes it difficult to know whether a move will be easy to learn. To be sure, the easiest way is to use gestures that are considered intuitive and natural by users.

4.2.5 Accuracy

The challenge with accuracy is to ensure the camera precisely detects, tracks and recognizes the gestures. According to Wachs et al. (2011, p. 62) because of characteristics such as hand shape, speed of movement, skin colour, hand size and the ambient light. Nevertheless, Leap motion currently allows for an extremely high level of accuracy, so this should not be forgotten, but the solutions exist.

4.2.6 Low mental load

Wachs et al. (2011, p.63) explains that to avoid mental load, "*gestures should be simple, temporally short, and natural*". Moreover, they add that iconic gestures could be a solution to avoid mental load as the meaning of these gestures is strongly linked with the intended actions. However, managers should beware of inheritance bias.

4.2.7 Intuitiveness

The main challenge with intuitiveness comes from the fact that this dimension is strongly related to the user's experience and cultural background (Wachs et al., 2011, p.63). A way to ensure intuitiveness is to conduct GES by using the Wizard-of-Oz technique according to Wachs et al. (2011, p.63). Concerning the cultural background, using the emblematic nature of gestures when classifying can be a good way to identify the gestures that are strongly linked with the user's culture.

4.2.8 Comfort

According to Wachs et al. (2011, p. 63), it is quite difficult to measure the stress induced by a gesture, specifically hand gestures. Managers must therefore find the best way to assess it by choosing either to conduct subjective questionnaires by directly asking the participants of the elicitation study, or by using electronic devices to measure muscle activity.

4.2.9 Lexicon size and multi-hand systems

The challenge lies in the fact that the robustness of the system has to be evaluated in order to know exactly how many gestures are still acceptable in order to ensure the proper functioning of the system (Wachs et al., 2011, p. 63)

4.2.10 "Come as you are"

This aspect was found to be of least importance in my exploratory study. According to Wachs et al (2011, p. 63) the main challenge with users not having to wear any sensors is that the vision algorithm must be able to correctly detect the characteristics of the users even though they evolved in different environments and have quite different physical appearances.

4.2.11 Reconfigurability

The challenge with reconfigurability is to find the right balance between allowing the user almost unlimited customisation possibilities and giving them enough flexibility to install the main components of the system as they wish (Wachs et al., 2011, p.64).

4.2.12 Gesture spotting and immersion syndrome

The main challenge, according to Wachs et al (2011, p.64) in order to avoid gestures being misinterpreted is that the system clearly detects the beginning as well as the end of the gesture via temporal landmarks. To solve this, they suggest combining the gesture with a voice or specific sound landmark. Another possible solution could be to switch gesture detection on and off, as shown in the GES by Chen et al. (2017), where one gesture is used to mean "gesture on" and another "gesture off".

4.3 Computing domains requirements

As for the requirements related to the computing domains explained above in terms of input (section 2.1.2.2), the table below (table 6) summarises the different levels of importance of the characteristics following the analysis with the Kano model. Managers should therefore determine the computing in which they wish to develop or use their system in order to know which features are indispensable.

	Productivity	**Media Consumption**	**Communication**	**Augmentation**
Must-be	Non-fatigue Accuracy Responsiveness	Intuitiveness	Responsiveness	Intuitiveness Responsiveness
One-dimensional	Speed	Responsiveness	Accuracy Speed (Flow)	Accuracy
Attractive	Intuitiveness	Accuracy	Non-fatigue Intuitiveness	Non-fatigue
Indifferent	/	Speed Non-fatigue	/	Speed (flow)

Table 6- Computing domains requirements in terms of input

4.4 Gesture representation in “quick learning guide”

McAweeney et al. (2018, pp. 5-8) conducted a second study when developing their taxonomy to represent gestures after GES, in which they analyse user-defined representations to better understand how end-users commonly apply their taxonomy elements in order to describe various aspects of the gesture. The study, which involved 30 participants, found that representation should generally articulate five aspects through graphic elements and design choices, these five aspects being Touch, Posture, Position, Motion and Time.

To communicate **Touch**, users predominantly used touchpoints (25 participants), but some (9) also used perspective changes. In addition, 8 participants used colour in the case of a hovering gesture “*for both touchpoints and literal shadows above a surface*” (McAweeney et al., 2018, p. 8). The study also revealed that if there was not a need to discern touchpoint and surface, in the case of a hovering gesture for instance, most users considered that using both was redundant, and therefore chose to eliminate either the drawn surface or touchpoints.

Regarding **Posture**, which corresponds to the necessary hand shape to perform a gesture correctly, 16 participants used perspective changes and colour for highlighting or clarifying the correct posture of the hand. Nevertheless, disrupting the coherence of perspective confused many of the participants performing the gesture, resulting in delayed processing and/or poor execution. In terms of gesture elements, the most chosen was text.

Concerning **Position**, participants tended to misunderstand three designs: fixed position; stop points relative to the rest of the body; and ending points also relative to the rest of the body. Consequently, participants used full or partial bodies, while still keeping unnecessary body parts to a minimum to avoid distracting the user from the place where the movement takes place, in order to add some context and point of view. However, in terms of the dimensions used, participants mostly used objects in the environmental context, twice as much as they used the body.

For the **Motion** aspect, participants tended, in half of the final representations, to use points or arrows on the joints where the bending and/or movement exactly occurred, to make the gesture more precise. In addition, participants agreed on the fact that clear indication must be provided to indicate start and end points of a movement, as well as the number of times a movement should be done in the case of two-sided arrows representing continuous motion.

Arrows were the most used gesture elements, followed by Ghost (when they represented lighter end points than start points).

Finally, the **Time** aspect was mainly represented by multi frames and/or numbers. Sometimes combined with arrows to communicate frames' order, numbers have remained the trend. Time aspect indeed represents "*the sequence of motion and hand posture in any given gesture*" (McAweeney et al., 2018, p. 7).

Therefore, although the above indications are only the point of view of a limited number of potential users of McAweeney et al.'s study, managers could clearly be inspired by them, or at least keep these results in mind in order to represent the gestures achievable with their interfaces, when classifying gestures in order to develop quick learning guide for their technologies.

4.5 Implications when classifying gestures post GES

In order to have a complete set of information, managers should use the taxonomy proposed in section 3.5. Indeed, this taxonomy allows to know the nature of the gesture (pointing, semaphoric, iconic, pantomimic, manipulative or emblematic); its form (static, dynamic or stroke); its binding (object-centric, world-dependent or world-independent); its temporality (discrete or continuous); its context (in-context or no-context); its complexity (simple or compound); the body part used (head, shoulder, arm, hand, torso, leg, foot, or whole body); its handedness (dominant, non-dominant or bi-manual); the hand shape used (flat hand, free hand, ASL shapes, etc.).); and finally the range of motion of the gesture (small, medium, large).

By using all of the above characteristics in his classification, a manager will therefore have enough information to analyse and develop a set of gestures with great diversity. And although care must be taken to ensure the robustness of the system to support such a variety of gestures, having such a high precision of gestures allows a better visibility to differentiate more easily between two gestures that would look alike. However, it is clear that having such a wide range of gestures could clearly cause problems in terms of memorability for users, and managers should therefore be cautious about the extent of the range of gestures that can be used with their interfaces.

Furthermore, such a wide range of classifications requires special attention during the study to ensure that all characteristics are taken into account and may require the user to spend more time proposing gestures during the study if the user is made aware of the variety of possibilities. However, making the user aware of the variety of gestures he can perform would, among other things, limit the inheritance bias as he would probably be less likely to repeat the gestures if he knows he has more possibilities.

In addition, whatever the method chosen to collect the gestures, managers should not forget to calculate the rate of agreement in order to be able to propose, following the classification of the gestures in their study, a set of gestures for which the majority of the participants have agreed. Nevertheless, they should not underestimate gesture inheritance bias and can propose a set of exclusive gestures, as Bostan et al. (2017) did in their study, if they realise from their classification that gestures "in-context" can in fact be used in the same context and should therefore be replaced.

One possible way to limit this inheritance bias, as well as the performance bias, proposed by Ruiz and Vogel (2015), is to use light constraints when eliciting whole-body gestures in order to limit arm fatigue. Ruiz and Vogel (2015, p. 3347) defined a soft constraint as something that physically limits certain motions and encourages others to be performed. In their study, in order to limit large arm movements, they forced the study participants to wear weighted wristbands. And this paid off as participants in their studies tended to propose alternative gestures that were less fatiguing to the arms, thus counteracting performance and inheritance bias (Ruiz & Vogel, 2015, p. 3349).

5. LIMITS

First of all, as a student of business engineering, and having no prior knowledge of the classification of gestures, I may have omitted one or other error when analysing the nature of the 571 gestures. However, errors were tried to be omitted as much as possible by following the different definitions of gestures from various trustworthy works summarised above.

In addition, another limitation is that despite having mixed several studies in the different fields, I was still strongly influenced by what I already knew, namely the entertainment field. The distribution between the different fields is therefore not the most optimal or representative.

A third limitation is that, although the proposed taxonomy seems relevant to me in view of the analysis of the different classifications, a deep analysis of the vision algorithm data was not realised as the needed computer knowledge to achieve this objective was beyond my competences.

Finally, as already mentioned, the number of participants in my online study is probably too small to be representative of the population. Therefore, the data selected can be used for exploration but an additional statistical study should be conducted in order to inspect the significance of the sample size.

6. CONCLUSION

To summarise, this work has shown that the main requirements for the development of 3D communication interfaces are: Price, Responsiveness, User adaptability and feedback, Learnability, Accuracy, Low mental load, Intuitiveness, Comfort, Lexicon size and multi-hands system, “Come as you are”, Reconfigurability, Interaction space, Gesture spotting and immersion syndrome, as well as Ubiquity and wearability. In addition, thanks to Yong (2014), the most important requirements could be classified by using the Kano model.

In addition, different methods of conducting GES have been listed, and although this is not an exhaustive list, the main ways of doing this are listed. These methods are as follows: the guessability method, the choice-based method, the “intuitive and ergonomic” method, as well as the Wizard-of-Oz technique. Furthermore, the main criteria to consider when conducting a GES are: ease of performing, intuitiveness, and memorability.

Then, after analysing the evolution of gesture types, and combining aspects of the taxonomies of Aigner et al. (2012), Vafaei (2013) and Hiyadi (2016), I could arrive at the following classification: **nature** (pointing, semaphoric, iconic, pantomimic, manipulative or emblematic); **form** (static, dynamic or stroke); **binding** (object-centric, world-dependent or world-independent); **temporality** (discrete or continuous); **context** (in-context or no-context); **complexity** (simple or compound); **body part** used (head, shoulder, arm, hand, torso, leg, foot, or whole body); **handedness** (dominant, non-dominant or bi-manual); **handshape** used (flat hand, free hand, ASL shapes, etc.).); and finally the **range of motion** (small, medium, large). This classification is the one that seemed to me the most optimal in order to have the most complete information possible on each gesture

Finally, the managerial implications behind the proposed taxonomy include being careful when conducting the gesture study not to overlook a factor, not to set aside Kano's analysis made in terms of input in order to know the requirements according to the chosen application domain, as well as to rely on McAweeney's (2018) recommendations in order to best represent gestures. In addition, managers also need to be mindful of inheritance and performance bias when conducting their elicitation studies and can do so with the help of soft constraints as explained by Ruiz and Vogel (2015).

7. BIBLIOGRAPHY

Aigner, R.,Wigdor, D., Benko, H., Haller, M., Lindlbauer, D., Ion, A., Zhao, S., & Koh, J. (2012). Understanding Mid-Air Hand Gestures: A Study of Human Preferences in Usage of Gesture Types for HCI.

Altmann, M. (2017). Designing Gestures for Window Management on Large High-Resolution Displays. Institute for Visualization and Interactive Systems University of Stuttgart. https://doi.org/10.18419/opus-9460

Bernsen, N. O., & Dybkjær, L. (2009). Multimodal Usability (Human Computer Interaction Series). Springer Publishing. 67-111.

Bostan, I., Buruk, O. T., Canat, M., Tezcan, M. O., Yurdakul, C., Göksun, T., & Özcan, O. (2017). Hands as a Controller: User Preferences for Hand Specific On-Skin Gestures. Proceedings of the 2017 Conference on Designing Interactive Systems, 1123–1134. https://doi.org/10.1145/3064663.3064766

Brito, I. V., Freire, E. O., Carvalho, E. A. N., & Molina, L. (2019). Analysis of Cross-Cultural Effect on Gesture-Based Human-Robot Interaction. International Journal of Mechanical Engineering and Robotics Research, 852–859. https://doi.org/10.18178/ijmerr.8.6.852-859

Cauchard, J. R., E, J. L., Zhai, K. Y., & Landay, J. A. (2015). Drone & me. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15, 361–365. https://doi.org/10.1145/2750858.2805823

Chan, E., Seyed, T., Stuerzlinger, W., Yang, X. D., & Maurer, F. (2016). User Elicitation on Single-hand Microgestures. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 3403–3414. https://doi.org/10.1145/2858036.2858589

Chen, Z., Ma, X., Peng, Z., Zhou, Y., Yao, M., Ma, Z., Wang, C., Gao, Z., & Shen, M. (2017). User-Defined Gestures for Gestural Interaction: Extending from Hands to Other Body Parts. International Journal of Human–Computer Interaction, 34(3), 1–46. https://doi.org/10.1080/10447318.2017.1342943

de Silva, S., Barlow, M., & Easton, A. (2013). Harnessing multi-user design and computation to devise archetypal whole-of-body gestures. Proceedings of the 25th Australian Computer-Human Interaction Conference on Augmentation, Application, Innovation, Collaboration - OzCHI '13, 85–94. https://doi.org/10.1145/2541016.2541020

Dim, N. K., Silpasuwanchai, C., Sarcar, S., & Ren, X. (2016). Designing Mid-Air TV Gestures for Blind People Using User- and Choice-Based Elicitation Approaches. Proceedings of the 2016 ACM Conference on Designing Interactive Systems, 204–214. https://doi.org/10.1145/2901790.2901834

Dong, H., Danesh, A., Figueroa, N., & el Saddik, A. (2015). An Elicitation Study on Gesture Preferences and Memorability Toward a Practical Hand-Gesture Vocabulary for Smart Televisions. IEEE Access, 3, 543–555. https://doi.org/10.1109/access.2015.2432679

Dong, H., Figueroa, N., & el Saddik, A. (2015). An Elicitation Study on Gesture Attitudes and Preferences Towards an Interactive Hand-Gesture Vocabulary. Proceedings of the 23rd ACM international conference on Multimedia, 999–1002. https://doi.org/10.1145/2733373.2806385

Felberbaum, Y., & Lanir, J. (2016). Step by Step: Investigating Foot Gesture Interaction. Proceedings of the International Working Conference on Advanced Visual Interfaces, 306–307. https://doi.org/10.1145/2909132.2926057

Feyereisen, P. (1994). Review of Hand and Mind: What Gestures Reveal about Thought, by D. McNeill. The American Journal of Psychology, 107(1), 149–155. https://doi.org/10.2307/1423299

Gentile, V., Fundarò, D., & Sorce, S. (2019). Elicitation and evaluation of zoom gestures for touchless interaction with desktop displays. Proceedings of the 8th ACM International Symposium on Pervasive Displays. https://doi.org/10.1145/3321335.3324934

Guerrero-García, J., González, C., & Pinto, D. (2017). Studying user-defined body gestures for navigating interactive maps. Proceedings of the XVIII International Conference on Human Computer Interaction, 1–4. https://doi.org/10.1145/3123818.3123851

Henze, N., Löcken, A., Boll, S., Hesselmann, T., & Pielot, M. (2010). Free-hand gestures for music playbackk: Deriving Gestures with a User-Centred Process. Proceedings of the 9th

International Conference on Mobile and Ubiquitous Multimedia - MUM ’10. https://doi.org/10.1145/1899475.1899491

Hiyadi, H. (2016, december). Reconnaissance 3D de gestes pour l’interaction homme-système (PhD Thesis). Université Paris-Saclay ; Université d’Evry-Val-d’Essonne ; Université Mohammed V de Rabat. 23-48. https://hal.archives-ouvertes.fr/tel-01804336

Joint Classification. (2022). Physiopedia. Accessed on 26 july 2022, at https://www.physio-pedia.com/Joint_Classification

Jurewicz, K. A., Neyens, D. M., Catchpole, K., & Reeves, S. T. (2018). Developing a 3D Gestural Interface for Anesthesia-Related Human-Computer Interaction Tasks Using Both Experts and Novices. Human Factors: The Journal of the Human Factors and Ergonomics Society, 60(7), 1–16. https://doi.org/10.1177/0018720818780544

Karam, M. & Schraefel, M.C. (2005). A Taxonomy of Gestures in Human Computer Interactions. University of Southampton. 1-45.

Kistler, F., & André, E. (2013). User-Defined Body Gestures for an Interactive Storytelling Scenario. Human-Computer Interaction – INTERACT 2013, 264–281. https://doi.org/10.1007/978-3-642-40480-1_17

Koutsabasis, P., & Domouzis, C. K. (2016). Mid-Air Browsing and Selection in Image Collections. Proceedings of the International Working Conference on Advanced Visual Interfaces, 21–27. https://doi.org/10.1145/2909132.2909248

Koutsabasis, P., & Vogiatzidakis, P. (2019). Empirical Research in Mid-Air Interaction : A Systematic Review. International Journal of Human–Computer Interaction, 35(18), 1747-1768. https://doi.org/10.1080/10447318.2019.1572352

Landragin, F. (2013). Man-machine dialogue : Design and challenges. John Wiley & Sons, Incorporated. 4-28

Lee, S. S., Chae, J., Kim, H., Lim, Y. K., & Lee, K. P. (2013). Towards more natural digital content manipulation via user freehand gestural interaction in a living room. Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 617–626. https://doi.org/10.1145/2493432.2493480

Legrenzi, C. (2016). Informatique, numérique et système d'information : définitions, périmètres, enjeux économiques. Vie & sciences de l'entreprise, N° 200(2), 49−76. https://doi.org/10.3917/vse.200.0049

Li, X., Guan, D., Zhang, J., Liu, X., Li, S., & Tong, H. (2019). Exploration of Ideal Interaction Scheme on Smart TV: Based on User Experience Research of Far-Field Speech and Mid-air Gesture Interaction. Design, User Experience, and Usability. User Experience in Advanced Technological Environments, 144–162. https://doi.org/10.1007/978-3-030-23541-3_12

Madapana, N., Gonzalez, G., Taneja, R., Rodgers, R., Zhang, L., & Wachs, J. (2019). Preference elicitation: Obtaining gestural guidelines for PACS in neurosurgery. International Journal of Medical Informatics, 130, 103934, 1–9. https://doi.org/10.1016/j.ijmedinf.2019.07.013

Manghisi, V. M., Uva, A. E., Fiorentino, M., Gattullo, M., Boccaccio, A., & Monno, G. (2018). 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, 32, 186–197. https://doi.org/10.1016/j.culher.2018.02.014

Magrofuoco, N., Perez-Medina, J. L., Roselli, P., Vanderdonckt, J., & Villarreal, S. (2019). Eliciting Contact-Based and Contactless Gestures With Radar-Based Sensors. *IEEE Access*, *7*, 176982–176997. https://doi.org/10.1109/access.2019.2951349

McAweeney, E., Zhang, H., & Nebeling, M. (2018). User-Driven Design Principles for Gesture Representations. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1-13. https://doi.org/10.1145/3173574.3174121

Min, H., Yun, J., & Geum, Y. (2018). Analyzing Dynamic Change in Customer Requirements : An Approach Using Review-Based Kano Analysis. Sustainability, 10(3), 746, 1-17. https://doi.org/10.3390/su10030746

Morris, M. R. (2012). Web on the wall: Insights from a Multimodal Interaction Elicitation Study. Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces - ITS '12, 95–104. https://doi.org/10.1145/2396636.2396651

Navarro-Newball, A. A., Moreno, I., Prakash, E., Arya, A., Contreras, V. E., Quiceno, V. A., Lozano, S., Mejìa, J. D., & Loaiza, D. F. (2015). Gesture based human motion and game principles to

aid understanding of science and cultural practices. Multimedia Tools and Applications, 75(19), 11699–11722. https://doi.org/10.1007/s11042-015-2667-5

Nebeling, M., Huber, A., Ott, D., & Norrie, M. C. (2014). Web on the Wall Reloaded: Implementation, Replication and Refinement of User-Defined Interaction Sets. Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces - ITS '14, 15–24. https://doi.org/10.1145/2669485.2669497

Nielsen, J. (1993, January 1). Response Time Limits: Article by Jakob Nielsen. Nielsen Norman Group. Retrieved 6 August 2022, from https://www.nngroup.com/articles/response-times-3-important-limits/

Nielsen, J. (2010, October 20). Website Response Times. Nielsen Norman Group. Retrieved 6 August 2022, from https://www.nngroup.com/articles/website-response-times/

Norton, J., Wingrave, C. A., & LaViola, J. J. (2010). Exploring strategies and guidelines for developing full body video game interfaces. Proceedings of the Fifth International Conference on the Foundations of Digital Games - FDG '10, 155–162. https://doi.org/10.1145/1822348.1822369

Obaid, M., Häring, M., Kistler, F., Bühling, R., & André, E. (2012). User-Defined Body Gestures for Navigational Control of a Humanoid Robot. Social Robotics, 367–377. https://doi.org/10.1007/978-3-642-34103-8_37

Obaid, M., Kistler, F., Kasparavičiūtė, G., Yantaç, A. E., & Fjeld, M. (2016). How would you gesture navigate a drone? A User-Centered Approach to Control a Drone. Proceedings of the 20th International Academic Mindtrek Conference. https://doi.org/10.1145/2994310.2994348

Pons, P., & Jaen, J. (2019). Interactive spaces for children: gesture elicitation for controlling ground mini-robots. Journal of Ambient Intelligence and Humanized Computing, 11(6), 2467–2488. https://doi.org/10.1007/s12652-019-01290-6

Ramis, S., Perales, F. J., Manresa-Yee, C., & Bibiloni, A. (2015). Usability Study of Gestures to Control a Smart-TV. Communications in Computer and Information Science, 135–146. https://doi.org/10.1007/978-3-319-22656-9_10

Range of Motion. (2022). Physiopedia. Accessed on 26 july 2022, at https://www.physio-pedia.com/Range_of_Motion

Rodriguez, I. B., & Marquardt, N. (2017). Gesture Elicitation Study on How to Opt-in & Opt-out from Interactions with Public Displays. Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces, 32–41. https://doi.org/10.1145/3132272.3134118

Rose, S., Clark, M., Samouel, P., & Hair, N. (2012). Online Customer Experience in e-Retailing: An empirical model of Antecedents and Outcomes. Journal of Retailing, 88(2), 308–322. https://doi.org/10.1016/j.jretai.2012.03.001

Rupprecht, D., Blum, R. & Bomsdorf, B. (2013). TOWARDS A GESTURE SET FOR A VIRTUAL TRY-ON. in Proceedings of IADIS International Conference Interfaces and Human Computer Interactions. IADIS Press. 273-277.

Salvador, R. A., & Naval, P. (2022). Towards a Feasible Hand Gesture Recognition System as Sterile Non-contact Interface in the Operating Room with 3D Convolutional Neural Network. Informatica, 46(1), 1-12. https://doi.org/10.31449/inf.v46i1.3442

Schiavo, G., Ferron, M., Mich, O., & Mana, N. (2016, 28 december). Wizard of Oz Studies with Older Adults : A Methodological Note. CINECA IRIS Institutional Research Information System. Retrieved on 19 July 2022 from https://cris.fbk.eu/handle/11582/308413.

Seuter, M., Macrillante, E. R., Bauer, G., & Kray, C. (2018). Running with drones: desired services and control gestures. Proceedings of the 30th Australian Conference on Computer-Human Interaction, 384–395. https://doi.org/10.1145/3292147.3292156

Silpasuwanchai, C., & Ren, X. (2015). Designing concurrent full-body gestures for intense gameplay. International Journal of Human-Computer Studies, 80, 1–13. https://doi.org/10.1016/j.ijhcs.2015.02.010

Smith, T. R., & Gilbert, J. E. (2018). Dancing to design. Proceedings of the 17th ACM Conference on Interaction Design and Children, 638–643. https://doi.org/10.1145/3202185.3210790

Sorathia, K., Jain, M., Amrit, M., Punekar, R. M., Srivastava, S., & Rajput., N. 2015. Gesture selection study for a maternal healthcare information system in rural Assam, India. J. Usability Studies 11, 1 (November 2015), 7–20.

Tracking | Leap Motion Controller | Ultraleap. (2022). Ultraleap. Retrieved 5 August 2022, from https://www.ultraleap.com/product/leap-motion-controller/

Vafaei, F. (2013, août). Taxonomy of Gestures in Human Computer Interaction (Thesis). North Dakota State University of Agriculture and Applied Science. http://hdl.handle.net/10365/23110

Vanderdonckt, J. (2021, 1 may). Conducting a Gesture Elicitation Study : How to Get the Best Gestures From People? [Diapositives]. SlideShare. https://www.slideshare.net/jeanvdd/conducting-a-gesture-elicitation-study-how-to-get-the-best-gestures-from-people

Vanderdonckt, J., Magrofuoco, N., Kieffer, S., Pérez, J., Rase, Y., Roselli, P., & Villarreal, S. (2019). Head and Shoulders Gestures: Exploring User-Defined Gestures with Upper Body. Design, User Experience, and Usability. User Experience in Advanced Technological Environments, 192–213. https://doi.org/10.1007/978-3-030-23541-3_15

Vatavu, R. D. (2012). User-defined gestures for free-hand TV control. Proceedings of the 10th European Conference on Interactive Tv and Video - EuroiTV ’12, 45–48. https://doi.org/10.1145/2325616.2325626

Villarreal-Narvaez, S., Vanderdonckt, J., Vatavu, R. D., & Wobbrock, J. O. (2020). A Systematic Review of Gesture Elicitation Studies: What Can We Learn from 216 Studies?. Proceedings of the 2020 ACM Designing Interactive Systems Conference. 855-872. https://doi.org/10.1145/3357236.3395511

Vogiatzidakis, P., & Koutsabasis, P. (2018). Gesture Elicitation Studies for Mid-Air Interaction: A Review. Multimodal Technologies and Interaction, 2(4), 65, 3-21. https://doi.org/10.3390/mti2040065

Vogiatzidakis, P., & Koutsabasis, P. (2019). Frame-Based Elicitation of Mid-Air Gestures for a Smart Home Device Ecosystem. Informatics, 6(2), 2–16. https://doi.org/10.3390/informatics6020023

Wachs, J. P., Kölsch, M., Stern, H., & Edan, Y. (2011). Vision-based hand-gesture applications. Communications of the ACM, 54(2), 60-71. https://doi.org/10.1145/1897816.1897838

Wittorf, M. L., & Jakobsen, M. R. (2016). Eliciting Mid-Air Gestures for Wall-Display Interaction. Proceedings of the 9th Nordic Conference on Human-Computer Interaction. https://doi.org/10.1145/2971485.2971503

Wu, H., Wang, J., & Zhang, X. (2014). User-centered gesture development in TV viewing environment. Multimedia Tools and Applications, 75(2), 733–760. https://doi.org/10.1007/s11042-014-2323-5

Wu, H., Yang, L., Fu, S., & Zhang, X. L. (2019). Beyond remote control: Exploring natural gesture inputs for smart TV systems. Journal of Ambient Intelligence and Smart Environments, 11(4), 335–354. https://doi.org/10.3233/ais-190528

Yong, K. F. (2014, january). Emerging Human-Computer Interaction Interfaces : A Categorizing Framework for General Computing (Thesis). Massachusetts Institute of Technology. http://hdl.handle.net/1721.1/90692

Zwass, V. (2020, November 2). information system. Encyclopedia Britannica. Retrieved July 5, 2022 from https://www.britannica.com/topic/information-system

8. APPENDIX

Appendix 1

Wu et al's (2014, pp. 744-745) gesture repartition representation

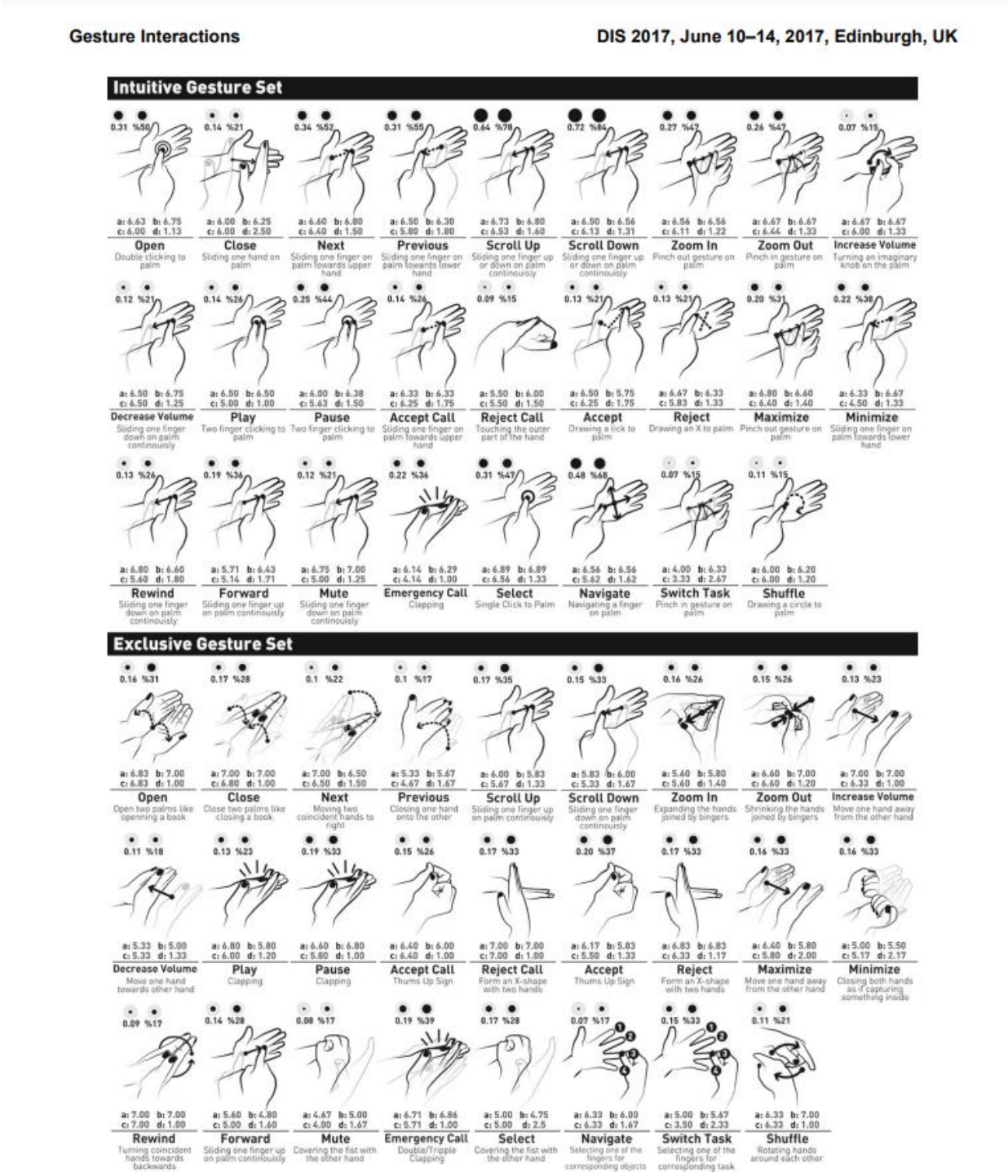
Table 1 Collected gestures

Function	Gesture description	Style	Frequency
Turn on TV	Both hands moving from the center middle to the outer left and right	Dynamic	7
	Single hand long pointing 5 s	Static	6
	Fist open	Dynamic	4
	Clap hands	Dynamic	2
	Finger snap	Dynamic	2
	An "OK" sign	Static	2
	Push forward	Dynamic	1
Turn off TV	Both hands moving from the outer left and right to the center middle	Dynamic	7
	Single hand long pointing 5 s	Static	6
	Fist close	Dynamic	3
	Clap hands	Dynamic	2
	A dynamic "pigtail" movement	Dynamic	2
	Finger snap	Dynamic	2
	Push forward	Dynamic	1
	An "OK" sign	Static	1
Turn to next channel	Swiping right	Dynamic	12
	Swiping left	Dynamic	11
	Thumb up	Static	1
Turn to last channel	Swiping left	Dynamic	12
	Swiping right	Dynamic	11
	Thumb down	Static	1
Increase volume	Swiping up	Dynamic	21
	Thumb pointing to the right	Static	2
	Both hands moving from the center middle to the outer left and right	Dynamic	1
Decrease volume	Swiping down	Dynamic	21
	Thumb pointing to the left	Static	2
	Both hands moving from the outer left and right to the center middle	Dynamic	1
Menu	Spreading hand	Static	5
	Finger snap	Dynamic	5
	Right hand moving from left to right	Dynamic	2
	Grasp	Dynamic	2
	Push forward	Dynamic	2
	Both hands moving from the center middle to the outer left and right	Dynamic	2
	Anticlockwise circle	Dynamic	2
	Fist	Static	1
	Double point	Dynamic	1
	Right hand moving from top to bottom	Dynamic	1
	Pull down	Dynamic	1
Confirm	An "OK" sign	Static	8
	Push forward	Dynamic	7

Function	Gesture description	Style	Frequency
	A dynamic "tick" movement	Dynamic	4
	Two fists collision	Dynamic	1
	A fist double click	Dynamic	1
	Fist	Static	1
	Thumb up	Static	1
	Grab	Dynamic	1
Mute	Hands forming as a "T" sign	Static	9
	Hands forming as a "*Kaddish*" sign	Static	5
	Waving back and forth	Dynamic	3
	Push forward	Dynamic	2
	A dynamic "pigtail" movement	Dynamic	2
	A dynamic "backslash" movement	Dynamic	1
	Clap hands	Dynamic	1
	Grab	Dynamic	1

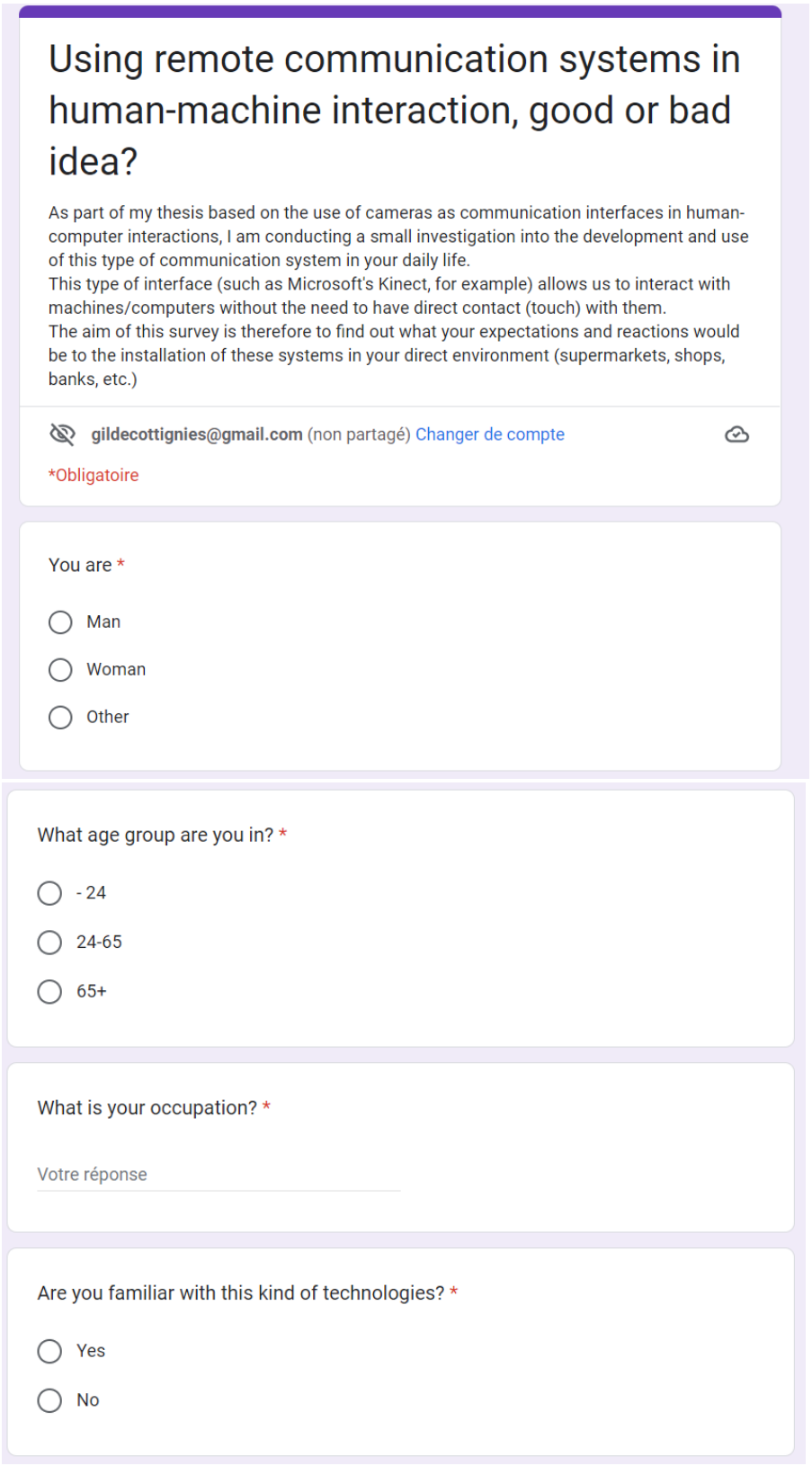
Appendix 2

Intuitive and Exclusive set representation from Bostan et al's (2017, p. 1127) GES



Appendix 3

Online google Form



Using remote communication systems in human-machine interaction, good or bad idea?

As part of my thesis based on the use of cameras as communication interfaces in human-computer interactions, I am conducting a small investigation into the development and use of this type of communication system in your daily life.
This type of interface (such as Microsoft's Kinect, for example) allows us to interact with machines/computers without the need to have direct contact (touch) with them.
The aim of this survey is therefore to find out what your expectations and reactions would be to the installation of these systems in your direct environment (supermarkets, shops, banks, etc.)

gildecottignies@gmail.com (non partagé) Changer de compte

*Obligatoire

You are *

- Man
- Woman
- Other

What age group are you in? *

- - 24
- 24-65
- 65+

What is your occupation? *

Votre réponse

Are you familiar with this kind of technologies? *

- Yes
- No

In which places (public or private) would you imagine being able to use this kind of technology?

Votre réponse

Would you be for or against the development of these kinds of technologies in your immediate environment? *

- ◯ For
- ◯ Against

Why are you in favour or against? *

Votre réponse

Some systems use sequences of gestures (making a sequence of at least 2 gestures) to perform a single task. Other systems use only simple gestures (gestures that do not need to be followed by another to be understood). Would you prefer to use a system programmed only with simple gestures; a system with simple and compound gestures; or is this not important to you? *

Votre réponse

Some systems use several identical gestures that vary in function depending on the context in which they are performed. Does this bother you? Why or why not? *

Votre réponse

What response time (in seconds) do you think is acceptable for the system to perform simple tasks (turning on the TV, turning up the volume, turning on the light, etc.)? *

Votre réponse

What response time (in seconds) do you think is acceptable for the system to perform average tasks (simple sorting (alphabetical order) in a file, downloading a small file, etc.)?

Votre réponse

What response time (in seconds) do you think is acceptable for the system to perform more complex tasks (e.g. asking to perform a complex calculation, downloading a large file, etc.)?

Votre réponse

How would you prefer the gestures to be presented to you in a "learning guide"?

- ◯ In a written way, as if someone was telling you how to do it.
- ◯ Visually, as if someone is showing you how to do it.

How would you prefer the gestures to be used to communicate with the machine to be represented visually (via a picture or diagram)?

- ◯ Visually with a "mirror" view
- ◯ Visually with a "sideways" view
- ◯ Visually with a "bird's eye" view (seen from above)

How do you rate your skills in communicating with machines/computers by gestures in order to control them? *

- ◯ I think I need help to learn how to use this type of interface (starting guide, tutorial, etc)
- ◯ I don't have any special skills, but I think I can manage to use this kind of interface on my own
- ◯ I feel I have enough knowledge of the gestures to use to interact with the interface
- ◯ I'm sure I'm more competent than most users of this kind of interface

How would you prefer the gestures to be used to communicate with the machine to be represented visually (via a picture or diagram)?

- ◯ Visually with a "mirror" view
- ◯ Visually with a "sideways" view
- ◯ Visually with a "bird's eye" view (seen from above)

How do you rate your skills in communicating with machines/computers by gestures in order to control them? *

- ◯ I think I need help to learn how to use this type of interface (starting guide, tutorial, etc)
- ◯ I don't have any special skills, but I think I can manage to use this kind of interface on my own
- ◯ I feel I have enough knowledge of the gestures to use to interact with the interface
- ◯ I'm sure I'm more competent than most users of this kind of interface

How would you rate your confidence in this type of communication interface? *

- ◯ Je considère que ce genre de système est fiable.
- ◯ I consider that I can rely on the quality of the supplier.
- ◯ I can have complete confidence, with no risk of uncertainty.
- ◯ I have blind faith in this type of system.

If you have already used this type of technology (Kinect, Leap motion, ...), what was your level of satisfaction?

- ◯ I was not satisfied and do not wish to use this type of system again.
- ◯ I am satisfied with my experience but no longer use this type of system.
- ◯ I am satisfied, and I am still using it (or plan to use it in the future)
- ◯ I am very satisfied and recommend to my friends and family to use this kind of system

How would you rate the control you feel you have over this type of communication interface? (if you have already used one)

- ◯ I have moderate control over the system and may make mistakes.
- ◯ I control the system fairly well but sometimes I make mistakeS
- ◯ I control the system correctly but sometimes I make a mistake.
- ◯ I have full control of the system and never make mistakes

How important do you think it is to customise such interfaces? *

- ◯ I would like to be able to program myself the gestures to perform the different actions without any assistance
- ◯ I would like to be offered a choice of gestures when configuring the system
- ◯ I am satisfied with the pre-recorded commands but would like to be able to adapt certain gestures
- ◯ I do not feel the need to modify the pre-recorded gestures

What importance do you attach to the body parts detected by the camera?

- ◯ I want to use only my hands and arms
- ◯ I want to be able to use my whole upper body
- ◯ I want to be able to use the lower part of my body
- ◯ I want to be able to use all parts of my body (from head to toe)

How important is the ease of execution of the gestures?

	1	2	3	4	5	6	7	
Unnecessary	◯	◯	◯	◯	◯	◯	◯	Essential

How important is the memorability of gestures to you?

	1	2	3	4	5	6	7	
Unnecessary	◯	◯	◯	◯	◯	◯	◯	Essential

How important is it to you to be comfortable when performing the actions?

	1	2	3	4	5	6	7	
Unnecessary	○	○	○	○	○	○	○	Essential

How important is the speed with which the system responds?

	1	2	3	4	5	6	7	
Unnecessary	○	○	○	○	○	○	○	Essential

How important is it for you to be able to "come as you are" (to be able to do everything through the camera without having to wear anything)?

	1	2	3	4	5	6	7	
Unnecessary	○	○	○	○	○	○	○	Essential

How important do you think it is that the system uses a natural and intuitive sign language (similar to human sign language)?

	1	2	3	4	5	6	7	
Unnecessary	○	○	○	○	○	○	○	Essential