# User-Defined Body Gestures for Navigational Control of a Humanoid Robot

Mohammad Obaid[1,2], Markus Häring[2], Felix Kistler[2], René Bühling[2], and Elisabeth André[2]

[1] HITLab New Zealand, University of Canterbury, Christchurch, New Zealand
[2] Augsburg University, Human Centered Multimedia, Augsburg, Germany

**Abstract.** This paper presents a study that allows users to define intuitive gestures to navigate a humanoid robot. For eleven navigational commands, 385 gestures, performed by 35 participants, were analyzed. The results of the study reveal user-defined gesture sets for both novice users and expert users. In addition, we present, a taxonomy of the user-defined gesture sets, agreement scores for the gesture sets, time performances of the gesture motions, and present implications to the design of the robot control, with a focus on recognition and user interfaces.

## 1 Introduction

Researchers are increasingly addressing the use of algorithms to recognize full body gestures and postures, in real time, to teleoperate and guide robots and hence enhance the user's natural experience and engagement with the robot, such as the work by [11][12]. The key to their approaches is to define intuitive and natural human-robot interaction using non-verbal communications, such as body gestures. Generally, most of the algorithms that use body gestures to control robots are based on gesture design paradigms that are defined by its developers. However, as the user is not involved in the process, the designed gestures may not be the most intuitive and may not represent their natural behavior. Recently, several researchers have addressed the same problem with the design of gesture based interaction methods in several other domains including surface computing [13] and public displays [5]. However, a user-defined set of gestures for the control of a humanoid robot has not been defined to this date.

In this paper, we present the design of a gesture set that is based on the user's natural behavior when controlling a robot. We collect data from both Technical[1] (T) and Non-Technical (NT) users when performing gesture motions to navigate a humanoid robot (Nao[2]). We contribute to the field of Human-Robot Interaction (HRI) the following: (1) the establishment of a user-defined gesture sets for both (T and NT users) to navigate a humanoid robot, (2) the analysis of qualitative and quantitative data that includes gesture taxonomy, performance data measures, observations, and subjective responses, and (3) an understanding of the implications for humanoid robot control using human gestures.

[1] We term a user experienced with robots and/or gesture tracking as *Technical*.
[2] `http://www.aldebaran-robotics.com`

## 2 Related Work

In this section, we present related literature and previous work on human gestures, designing gestures, and gesture controlled robots.

**Human Gesture Categories:** Researchers have conducted a vast number of studies to understand gestural interactions between individuals and how gestures can be categorized based on the information communicated. There is no universal categorization standard for body gestures and postures, however, researchers used different taxonomies for categorization. Efron [1] was one of the first to classify gestures into five categories: physiographics, kinetographics, ideographics, deictics, and batons. McNeill [6] presented six types of gestures: adaptor, beat, emblemic, deictic, iconic, and metaphoric gestures. Moreover, McNeil [7] defined four phases that construct a gesture: preparation, stroke, hold and retraction. The preparation is the phase that brings the body from its rest to a position that is suitable for executing the gesture. The stroke phase is the real information contained in the gesture, while the retraction is the phase where the body goes to its rest position again. Hold, on the other hand, is the temporal duration of the stroke phase. In this paper, we use the phases defined by McNeill.

**Designing Gestural Input:** The basic rule when designing an interface is to initially define the needs of its users and gestural interfaces are no exception [8]. Therefore, several domain areas employ the design of appropriate gestures for a system by allowing users to intuitively define how they would use it. Recently, the work presented by Wobbrock et al. [13] described the design of appropriate gestures for surface tabletop interfaces. They define gestures by employing non-technical users to observe the effect of a gesture and then asked them perform a gesture to match its cause. The work by Wobbrock et al. was a motive for many researchers to follow a similar design paradigm in their domain field. For example, Kray et al. [4] identified user-defined gestures that can be used to communicate a mobile phone with public display, tabletops, and other devices. Kurdyukova et al. [5] presented a study for identifying a user-defined set to transfer data from an iPad in a multi-display environment. In this research, we follow a similar approach to Wobbrock et al., with a focus on navigational gestural control for humanoid robots.

**Gesture Controlled Robots:** The fact that humanoid robots are machines that look like humans and preserve some human functionalities has motivated researchers to look for intuitive interaction ways that are similar to the human-human communications. While some work follows multimodal approaches, mostly combining speech with gesture commands [11], other work efforts are put towards controlling robots using pointing gestures [10], but such methods are limited to a certain range of commands. Moreover, Hu et al. [2] developed simple hand gestures for robot navigational actions, while, the recent work of Konda et al. [3] employ full body postures to navigate their robots.

Previous work, in this field, relied on the developers of the system to define commands and gestural instructions, however, none have exhibited how users would like to control a humanoid robot intuitively, which is the novel part presented in this paper.

## 3 User Defined Gestures to Control Humanoid Robots

The main objective of this study is to define a set of control body gestures derived from the users' actions when intuitively instructing a humanoid robot. In particular, in this study, we focus on navigational control of the humanoid robot Nao. We use eleven actions (*Move Forward, Move Backward, Move Left, Move Right, Turn Left, Turn Right, Stop, Speed Up, Slow Down, Stand Up, Sit Down*) for which users, of the presented study, chose gestures. The motions of all navigational actions are implemented from the perspective of the robot using the built in motion module of the Nao system (Academic Edition V3.2).

**Experimental Setup:** To define a set of intuitive gestures to control a humanoid robot, we consider two types of user groups, Technical (T) and Non-Technical (NT): The first are users that have some experience with humanoid robots and are aware of gesture tracking systems (such as Microsoft Kinect). The second are users who have no sound knowledge of such technologies. We consider the two groups as it is apparent when a user is aware of the limitation of the technologies they can define their gestures based on those limitations; hence, including the two groups (T and NT) allows system designers to consider the characteristics of both groups.

We elicit preformed gestural actions from 35 participants (17 T, 18 NT), all from Germany. Initially, we asked participants, on a 5-point Likert scale (ranging from one to five), about their experience with the Microsoft Kinect and with a humanoid robot. The 17 T participants (six female, eleven male) have an average experience with MS Kinect=2.71 and with a humanoid robot=2.41. The 17 T participants have an average age of 29 ($SD$ = 5.2) and are mainly from the Computer Science background. On the other hand, the 18 NT participants (ten female, eight male) have an average experience with MS Kinect=1.11 and with a humanoid robot=1.06. Most of the 18 NT participants are students from several disciplines, such as education, languages or economics, and have an average age of 27 ($SD = 7.8$). All participants except one were right-handed.

**Apparatus**: The experiment is arranged in a room with about 3 meters width and 6.5 meter depth. The room is equipped with a 50 inch plasma display and two cameras. The first camera records the front view of the user, while the other camera is setup as a side camera. The user has a designated region that he/she is allowed to freely move in during the study. This region is defined from the user's initial position and a distance of about 1 meter around that point. The humanoid robot, Nao, is placed to be facing the user at about 2 meters away from them.

**Table 1.** Taxonomy for full body gestures used to control a humanoid robot based on 385 gestures

| **Taxonomy of Full Body Gestures for Controlling a Humanoid Robot** | | |
|---|---|---|
| Form | static gesture | A static body gesture is held after a preparation phase. |
| | dynamic gesture | The gesture contains movement of one or more body parts during the stroke phase. |
| Body Parts | one hand | The gesture is performed with one hand. |
| | two hands | The gesture is performed with two hands. |
| | full body | The gesture is performed with at least one other body part than the hands. |
| View-Point | independent | The gesture is independent from the view point. |
| | user-centric | The gesture is performed from the user's point of view. |
| | robot-centric | The gesture is performed from the robot's point of view. |
| Nature | deictic | The gesture is indicating a position or direction. |
| | iconic | The gesture visually depicts an icon. |
| | miming | The used gesture is equal to the meant action. |

**Procedure**: At the beginning of the experiment, each participant is given a description of the study and are told to stay within their designated region in the room. The following are the steps each participant is asked to follow: (1) on the screen, watch a video that demonstrates how Nao performs one of the navigational actions. (2) Upon the completion of the video, perform a gesture that can command Nao to repeat the demonstrated action. (3) Watch Nao performing the corresponding action (this is remotely activated by an instructor). (4) Answer a questionnaire corresponding to the action.

The eleven actions are presented to each participant in a randomized order. For the actions *Speed up, Slow down* and *Stop*, Nao will be in motion when the gesture is to be preformed by the participant. In this case, participants are asked to state when they are ready, after watching the video on the screen, and Nao is immediately activated by the instructor. Subjective and objective measures are explained further in Section 4.

## 4 Results

The results of our study presents a gesture taxonomy, a user-defined gesture set, performance data measures, qualitative observations, and subjective responses.

**Gesture Taxonomy.** We manually classify all gestures according to four dimensions: *form*, *(involved) body parts*, *view-point*, and *nature*. Each dimension consists of multiple items, shown in Table 1. They are partly based on the Taxonomy used by Wobbrock et al. [13] and adapted to match full body gestures. Moreover, *nature* was inspired by gesture categories defined by Salem et al. [9].

*Form* in our sense distinguishes between static and dynamic gestures (without and with movement respectively). Static gestures have a preparation phase at

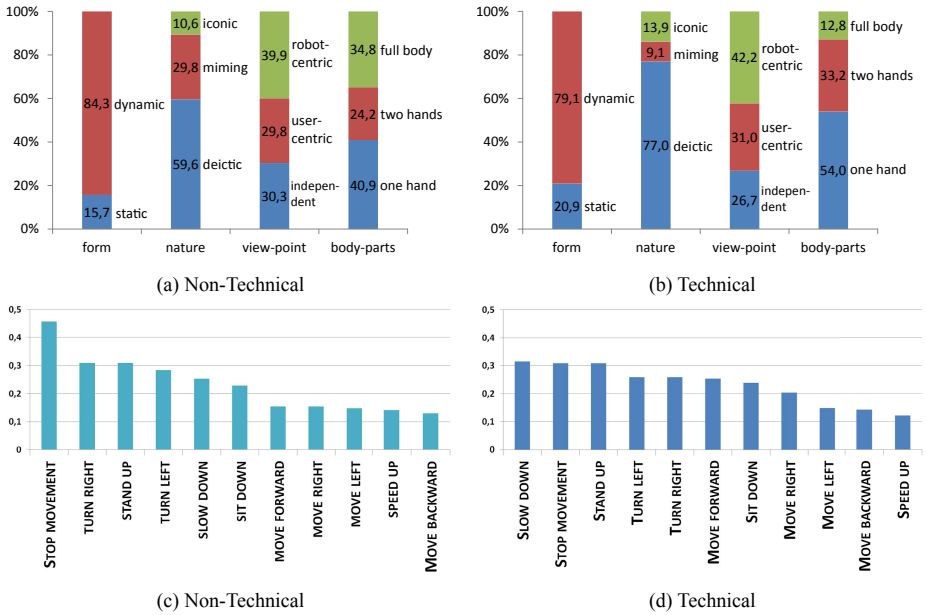


**Fig. 1.** Taxonomy distribution (a and b) and gesture agreement levels (c and d) for technical and non-technical users

the beginning, in which the user moves into the gesture space, but the core part of gesture is after the preparation phase. Therefore, the gesture is kept for a certain amount of time before the user releases it again in the retraction phase. In opposite, dynamic gestures have a clear stroke phase including the movement of body specific parts between the preparation and retraction phases.

The *body parts* dimension is quite self-explaining. It distinguishes between one hand, and two hand gestures, as well as full body gestures that involve at least one other body part.

The *view-point* dimension can be explained best with pointing gestures in a scenario where the robot is facing the user. Thus, a user-centric view-point means that when the user is pointing to his/her right, the robot should move in the pointing direction and, therefore, to the left from the robot's view. The opposite is a robot-centric view-point, i.e. when the user is pointing to his/her right, the robot moves in opposite to the pointing direction (to the right from the robot's view). Other gestures are view-point independent, for example, an open front-facing hand for stop which does not include any directional information.

The *nature* of our gesture is divided in three categories: The most common gestures we found for HRI are deictic gestures, that indicate a position or direction. These gestures can be either static, e.g. pointing to the right, or dynamic, e.g. waving to the right. They can be performed with one hand, two hands, or even other body parts, e.g. tilting the head. They can be performed from a user-centric or robot-centric view-point. Iconic gestures are visual depictions, e.g. an

open front-facing hand for stop, or drawing a circle in the air for turning. Miming gestures realize the idea that the user shows the robot how to perform the action by actually performing it, e.g. if the action is sitting down, the user actually sits down. Depending on the view-point, miming gestures can be mirrored as well.

Fig. 1 depicts the taxonomy distributions for T and NT users. The two most visible differences between the two kinds of users can be seen in the *nature* dimension ($\chi^2$ (2) = 26.36, $p < 0.001$) and the *involved body parts* dimension ($\chi^2$ (2) = 25.46, $p < 0.001$). While T users clearly prefer deictic gestures and mainly use their hands for gesturing, NT users more often use full body and miming gestures. Therefore, one can say that T users prefer more abstract and less exhausting gestures. This is emphasized by the fact that the T users also tend to use more static postures than the NT, however, we found no significant differences for the *form* dimension ($\chi^2$ (1) = 1.75, $p = 0.186$).

**A User-defined Gesture Set:** The gestural data collected from the participants of the study, to control the humanoid robot Nao, is used to define a set of user-defined gestures that can be used for navigations. The process of selecting a suitable gesture for a control action is as follows: (1) For each control action $t$ we identify a set $P_t$ that contains all proposed gestures. (2) The proposed gestures in $P_t$ are then grouped into subsets of identical gestures $P_{i_{1..N}}$, where $i$ is a subset that contains identical gestures and $N$ is the total number of identified subsets. (3) The representative gesture for action $t$ is identified by selecting the subset $P_i$ with the largest size, i.e. $MAX(P_i)$.

Fig. 2 depicts the representing gestures for the eleven actions for both T and NT users. In some cases, two representative gestures are present for one action as there were two large size gestural subsets ($P_i$) with an equal number of identical gestures, e.g. Action 1 for NT.

To further evaluate the degree of agreement among participants towards the selected user-defined sets, we employ a process that computes an agreement score[3] based on the work defined and used by Wobbrock et al [13]. An agreement score $S_t$ corresponding to a selected user-defined gesture for action $t$ is represented by a number in the range $[0, 1]$ that defines the general agreement among participants. The results of evaluating the degree of agreement for the eleven control actions of our study are presented in Fig. 1 (c) and (d). The overall agreement levels for the T and NT participants are the same, $S = 0.23$.

**Gestural Phases and Timing:** The video recordings of all participants, from the camera videotaping the frontal view of the user, were annotated using the ELan annotation tools[4]. The annotations segmented each video into 11 actions and each action into 4 phases (Start-up, Preparation, Stroke, and Retraction). The start-up phase represents the time it takes the participants to start their gestural instruction, after watching the action on the screen. While the others are the times for the gestural phases defined by McNeill [7]. Using the annotation tool, the

[3] For the equation refer to [13].

[4] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands (`http://www.lat-mpi.eu/tools/elan/`).

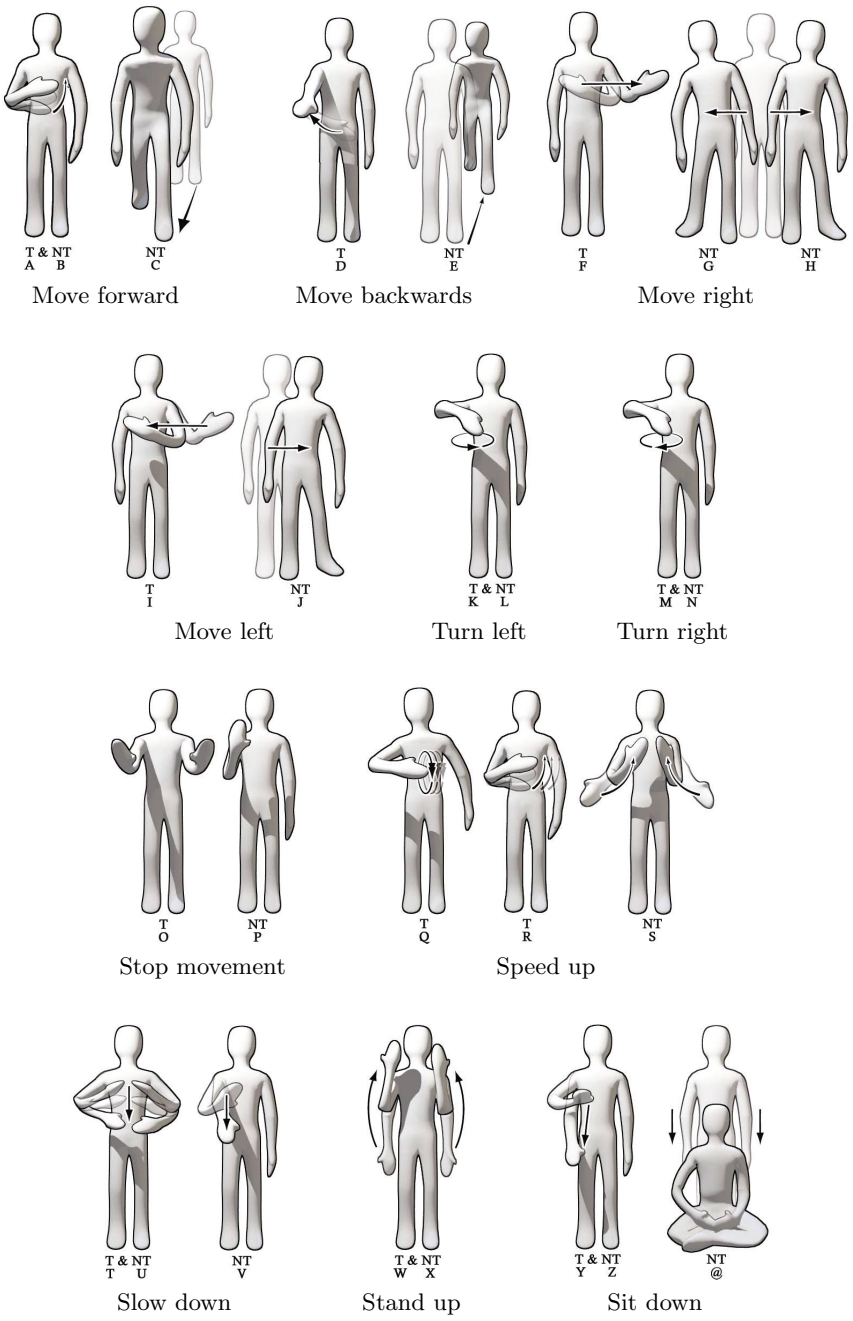


**Fig. 2.** User-defined gesture sets for the technical (T) and non-technical (NT) participants to navigate a humanoid robot

times for the 4 phases are extracted for the 11 actions of each participant. Table 2 shows the average times (for T and NT) for each of the phases of each gesture representing an action, and corresponds to Fig. 2.

**Subjective Ratings:** After each action, participants are asked to rate the *goodness* and *easiness* of their performed gesture on 7-point Likert scales. The results reveal that the *goodness* of the gestures and the *easiness* to think of them correlated significantly for the T group ($r = 0.54, p < 0.01$) as well as for the NT group ($r = 0.40, p < 0.01$). As expected, gestures that are considered as good matches for an action are usually easy to think of and to produce. Beside the direct correlation between *goodness* and *easiness*, we also checked for their correlation with the level of agreement and the timings (especially the *StartUp* and *Stroke* phase) but nothing significant could be found.

**Table 2.** Time in seconds (Mean, SD) for each of the four phases (**St**art-up, **Pr**eparation, **S**tro**k**e, and **Re**traction) for T and NT Participants. Labels correspond to the user-defined gesture sets illustrated in Fig. 2.

| **T** | **St** | **Pr** | **Sk** | **Re** | **T** | **St** | **Pr** | **Sk** | **Re** |
|---|---|---|---|---|---|---|---|---|---|
| A | 2.89, 1.71 | 0.29, 0.05 | 1.84, 1.21 | 0.80, 0.44 | D | 2.26, 0.91 | 0.40, 0.23 | 1.61, 0.47 | 1.12, 0.79 |
| F | 3.00, 1.43 | 0.42, 0.39 | 2.04, 1.91 | 0.84, 0.53 | I | 1.78 0.60 | 0.70, 0.62 | 2.37, 2.34 | 0.80, 0.24 |
| K | 2.59, 1.43 | 0.59, 0.57 | 2.28, 1.41 | 1.36, 0.86 | M | 2.48, 1.10 | 0.43, 0.18 | 2.19, 0.97 | 0.75, 0.40 |
| O | 2.76, 1.26 | 0.32, 0.12 | 2.14, 0.46 | 0.57, 0.25 | Q | 2.07, 1.54 | 0.24, 0.07 | 3.05, 1.51 | 0.74, 0.44 |
| R | 2.27, 1.04 | 0.25, 0.05 | 1.60, 0.11 | 0.63, 0.10 | T | 3.12, 2.36 | 0.37, 0.31 | 3.72, 1.76 | 1.25, 0.83 |
| W | 2.71, 0.92 | 0.57, 0.38 | 2.00, 1.21 | 0.95, 0.22 | Y | 1.69, 0.65 | 0.37, 0.18 | 2.11, 2.29 | 0.76, 0.21 |
| **NT** | **St** | **Pr** | **Sk** | **Re** | **NT** | **St** | **Pr** | **Sk** | **Re** |
| B | 2.12, 1.52 | 0.15, 0.30 | 4.84, 1.87 | 0.72, 1.44 | C | 2.04, 0.65 | 0.54, 0.34 | 2.02, 1.37 | 0.63, 0.17 |
| E | 2.19, 1.57 | 0.27, 0.54, | 4.22, 2.18 | N/A | G | 1.00, 0.52 | 0.17, 0.33 | 5.10, 1.65 | N/A |
| H | 2.39, 0.69 | 0.24, 0.28 | 2.85, 1.38 | 0.50, 1.01 | J | 1.44, 0.62 | 0.09, 0.19 | 4.59, 1.54 | N/A |
| L | 2.32, 2.18 | 0.73, 0.76 | 3.91, 2.45 | 0.70, 0.69 | N | 1.10, 0.48 | 0.45, 0.21 | 2.83, 1.60 | 0.73, 0.31 |
| P | 2.48, 1.26 | 0.24, 0.13 | 2.52, 0.92 | 1.09, 1.09 | S | 2.67, 1.36 | 0.28, 0.13 | 2.86, 2.03 | 0.95, 0.28 |
| U | 2.81, 1.17 | 1.06, 1.46 | 3.09, 1.23 | 1.16, 0.47 | V | 2.92, 2.24 | 0.89, 1.50 | 2.56, 1.73 | 1.35, 0.80 |
| X | 1.93, 0.94 | 0.72, 0.72 | 3.67, 1.92 | 0.38, 0.38 | Z | 1.35, 0.27 | 0.35, 0.11 | 1.81, 1.05 | 0.58, 0.12 |
| @ | 2.46, 1.58 | 0.99, 0.66 | 6.92, 2.32 | 1.52, 2.08 | | | | | |

## 5 Discussion

In this section, the implication of the results for the user-defined set of gestures to navigate a humanoid robot are discussed for both gesture recognition and user interfaces.

**Implication for Gesture Recognition:** The most user-defined gestures for navigational control of a humanoid robot are deictic gestures, which indicate a position or direction. Therefore, the main focus of the gesture recognition should lay on these type of gestures. However, we notice that the gesture view-point may vary especially in these cases. This poses a great challenge for the gesture recognition: if mirrored gestures should be allowed, how does the robot know if it should move to the left-hand or right-hand side, when the user is pointing to his/her right? A solution could be to offer different modes for the navigational control: one in robot-view and one in user-view. Nevertheless, the interaction

designer should think carefully of which gestures are influenced by the control mode. For example, gestures for linear movements are usually all influenced depending on the chosen view-point, while gestures for rotating the robot remain the same. Another interesting point is, that one-hand gestures are still the most important ones, however two-hand gestures are also used quite often, and NT users also performed quite a lot of gestures that involve other body parts. The usage of the second hand mostly results in symmetrical gestures, for which the information from the second hand is, more or less, redundant, but could be used to increase the confidence of a recognition system. The use of full body gestures raises a different issue: they can only be included when implementing additional gesture recognizers, and in opposite to the hand gestures, they really need the full body tracking information which justifies the usage of a depth sensor with corresponding tracking technology. Users generally preformed dynamic gestures, therefore, simple posture recognition would often be not enough. Moreover, the usual statically labeled pointing gesture should not be optimized for a certain amount of dwell-time as a lot of users included a single or repeated waving motion into pointing to indicate direction.

**Implication for User Interfaces:** In general, participants (both T and NT) had an affirmative response toward using freehand and full body gestures to navigate a robot. Throughout the study, an informal feedback was given by participants that include how easy it was to control a robot that way and how it can allow them to create a more realistic environment in controlling a humanoid robot. In some cases, the robot is described like a companion (or a pet) that can be ordered to move around using gestures. Nevertheless, it is notable that participants tend to talk to the robot during the study, even though they are aware that the robot does not respond to spoken commands. When participants are asked about why they gave a spoken command, 22 (11 T and 11 NT), or 63% of the total number of participants, stated that they would prefer a combination of speech and gesture commands, i.e. a multimodal interface, to control the robot. Moreover, several of the T participants indicated that they would prefer to make the robot stop with a speech command, while, NT participants would prefer to make the robot turn according to spoken instructions. Two of the NT participants also stated that they would prefer to control the robot only with speech commands than gestures, while none of the T participants would prefer it this way. In addition, it is apparent that participants are quick to respond to their task and produce gestures that correspond to the robot's navigational actions, where the overall average time, in seconds, for starting up a gesture after watching the action on screen is 2.25 (SD=0.57).

Moreover, several participants in the T group were worried that the recognition system will misclassify what their gesture was and the robot would do an unexpected action. On the other hand, participants in the NT groups expressed that they were worried that the robot will misunderstand their gestural command and perform a different action. This explains why a large number of participants would also give spoken commands in combination with their gestures; in addition, the importance of a reliable and robust gesture recognition system is vital in this case.

## 6 Conclusion and Future Work

In this paper, we have presented the results of a study to produce a user-defined gesture set to navigate a humanoid robot intuitively. The presented results are based on collecting data from two groups of users: technology aware users (i.e. gesture recognition and robots), and non-experienced users. The analysis of the data revealed a user defined-gesture set to control a humanoid robot. In addition, we presented (1) a taxonomy of the human-robot navigational gestures, (2) user agreement scores for each of the gestures representing a navigational commands, (3) time performances of the gesture motions, and (4) design implications for both recognition and user interfaces.

In the presented study, we focused on navigational commands, however, a humanoid robot can do more functions that can be also investigated in future work. In addition, the study revealed that a combination between gesture and speech commands is important and will be investigated in future work. Finally, we plan to implement the recognition of the user-defined gesture set in our open source Full Body Interaction Framework (FUBI)[5] and validate its functionality.

**Acknowledgments.** This work was partially funded by the European Commission within the 7th Framework Program under grant agreement eCUTE (FP7-ICT-257666).

## References

1. Efron, D.: Gesture and Environment. King's Crown Press, Morningside Heights, New York (1941)
2. Hu, C., Meng, M., Liu, P., Wang, X.: Visual gesture recognition for human-machine interface of robot teleoperation. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003), vol. 2, pp. 1560–1565 (October 2003)
3. Konda, K.R., Königs, A., Schulz, H., Schulz, D.: Real time interaction with mobile robots using hand gestures. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI 2012, pp. 177–178. ACM, New York (2012)
4. Kray, C., Nesbitt, D., Dawson, J., Rohs, M.: User-defined gestures for connecting mobile phones, public displays, and tabletops. In: Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services, MobileHCI 2010, pp. 239–248. ACM, New York (2010)
5. Kurdyukova, E., Redlin, M., André, E.: Studying user-defined ipad gestures for interaction in multi-display environment. In: International Conference on Intelligent User Interfaces, pp. 1–6 (2012)
6. McNeill, D.: So you think gestures are nonverbal? Psychological Review 92(3), 350–371 (1985)
7. McNeill, D.: Head and Mind: What Gestures Reveal About Thought. University of Chicago Press, Chicago (1992)

[5] `http://www.hcm-lab.de/fubi.html`

8. Saffer, D.: Designing Gestural Interfaces. O'Reilly Media, Sebastopol (2009)
9. Salem, M., Rohlfing, K., Kopp, S., Joublin, F.: A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In: RO-MAN, 2011 IEEE, July 31-August 3, pp. 247–252 (2011)
10. Sato, E., Yamaguchi, T., Harashima, F.: Natural interface using pointing behavior for human ndash;robot gestural interaction. IEEE Transactions on Industrial Electronics 54(2), 1105–1112 (2007)
11. Stiefelhagen, R., Fugen, C., Gieselmann, R., Holzapfel, H., Nickel, K., Waibel, A.: Natural human-robot interaction using speech, head pose and gestures. In: Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), September 28-October 2, vol. 3, pp. 2422–2427 (2004)
12. Suay, H.B., Chernova, S.: Humanoid robot control using depth camera. In: Proceedings of the 6th International Conference on Human-Robot Interaction, HRI 2011, pp. 401–402. ACM, New York (2011)
13. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 1083–1092. ACM, New York (2009)